Cyclic Causal Models with Discrete Variables: Markov Chain Equilibrium Semantics and Sample Ordering

David Poole* **and Mark Crowley†**
* Computer Science, University of British Columbia, http://cs.ubc.ca/~poole/
† Electrical Engineering & Computer Science, Oregon State University, http://markcrowley.ca

Abstract

We analyze the foundations of cyclic causal models for discrete variables, and compare structural equation models (SEMs) to an alternative semantics as the equilibrium (stationary) distribution of a Markov chain. We show under general conditions, discrete cyclic SEMs cannot have independent noise; even in the simplest case, cyclic structural equation models imply constraints on the noise. We give a formalization of an alternative Markov chain equilibrium semantics which requires not only the causal graph, but also a sample order. We show how the resulting equilibrium is a function of the sample ordering, both theoretically and empirically.

1 Introduction

Pearl [2009] advocates structural equation models (SEMs) as a representation for causality. A structural equation model consists of a deterministic function for each variable in terms of other variables and (independent) exogenous "noise" inputs. In this paper we analyze SEMs for cyclic models. A modal logic for SEMs was presented by Halpern [2000].

An alternative to the SEMs is an equilibrium model [Strotz and Wold, 1960], where the causes of each variable form a transition model of a Markov chain, and we are interested in the equilibrium distribution of this Markov chain. Strotz [1960] describes the contribution of that paper as:

> *If a causal interpretation of an interdependent system is possible it is to be provided in terms of a recursive system. The interdependent system is then either an approximation to the recursive system or a description of its equilibrium state.*

The SEM and the equilibrium structure of Iwasaki and Simon [1994] can be seen as an equilibrium where the *values* of the variables are invariant. In the Markov chain semantics the equilibrium is on the *distribution* of the variables.

In this paper we analyze the case where there are probabilistic effects of interventions and the graph of causal dependency can contain cycles. Note that Strotz and Wold [1960] and Pearl [2009] explicitly consider probabilities, using what Strotz and Wold [1960] call stochastic variables. Iwasaki and Simon [1994] do not explicitly include uncertainty.

There are two main reasons why we are interested in cyclic models:

- We may have information about the effect of interventions that is not acyclic, and we may need to incorporate this information into a model that also includes non-interventional data.

- There are models where there is no natural acyclic order. This occurs, for example, in spatial domains and in relational domains. In spatial domains each location in space may depend on its neighbours, for example, Crowley and Poole [2011] describe an application where in the policy of an MDP, the action at each location depends on the actions at other neighbouring locations. These models are causal because the actions are meant to be carried out by people in the field. Domingos *et al.* [2008] give an example where friends of friends are friends; they give an undirected model, but a causal model would need to be cyclic. A causal model is appropriate here because we would expect a causal model to be stable under changing populations.

There is a literature on conditional specifications of distributions where the graph induced by the conditional probabilities is cyclic [Heckerman *et al.*, 2000; Arnold *et al.*, 2001; Gelman, 2004; Neville and Jensen, 2007]. In each of these, (approximate) conditional probabilities are input or learned. In the motivating examples above, the causal probabilities do not correspond to conditional probabilities, but to the probabilistic effect of interventions.

In this paper we show limitations of standard SEM models for representing general cyclic causal models, and argue that an equilibrium semantics often makes more sense. The equilibrium, however, depends on a sample ordering; even if we know all of the direct causal effects, it is not enough information to compute the equilibrium distribution. We show cases where we can bound how much different orderings influence the distribution, and cases where the structure of the cyclic model induces classes of orderings which result in the same distributions.

2 Cyclic Causal Models and SEMs

In this paper, we only consider discrete variables. All examples use Boolean variables. We write variables in upper case and values in lower case. For a Boolean variable, X we write the instantiation $X = true$ as the lower-case of the variable, x, and $X = false$ as $\neg x$.

We use the do notation of Pearl [2009], where $P(x \mid do(y))$ means the probability of x after an intervention to make y true.

A causal model is a directed graph with variables as the nodes. One of the properties of the causal theories of Pearl is that causal models are sufficient to predict all combinations of interventions (including the case of no interventions). For each variable X, with parents π_X, and for each combination of values, $\overline{v}$, to π_X, the probabilities $P(X \mid do(\pi_X = \overline{v}))$ specify the model.

Example 1. Consider the simple cyclic causal model with two Boolean variables A and B, each dependent on the other. The causal model can be defined in terms of 4 parameters:

$$p_1 = P(a \mid do(b))$$
$$p_2 = P(a \mid do(\neg b))$$
$$p_3 = P(b \mid do(a))$$
$$p_4 = P(b \mid do(\neg a))$$

These probabilities can be obtained by intervening on one of the variables and observing the effect on the other variable in randomized controlled experiments.

A structural equation model M for Boolean variables is a set of sentences in propositional logic. There is one sentence for each variable in the model. The sentence describes how the variable depends on other variables in the model and exogenous (noise) variables. The exogenous variables have (independent) probability distributions over them. An intervention $X = v$ on a variable X replaces the sentence for X with a sentence that specifies that X has value v. We write the resulting model as $M_{X=v}$. An intervention on multiple variables is equivalent to an intervention on each variable in turn.

Example 2. The causal model of Example 1 can be represented as a structural equation model M:

$$a \leftrightarrow (b \wedge u_1) \vee (\neg b \wedge u_2) \quad (1)$$
$$b \leftrightarrow (a \wedge u_3) \vee (\neg a \wedge u_4) \quad (2)$$

where each U_i is an exogenous Boolean variable with $P(u_i) = p_i$. We say U_i is a noise variable.

If we intervene to make $A = false$, we replace (1) with $\neg a$ forming a model $M_{\neg a}$, containing $\neg a$ and equation (2).

Let M be a structural equation model, u a set of instantiations to exogenous variables, and α a propositional formula. $M, u \models \alpha$ means α is true in all models where M and u are true.

Example 3. In model M of Example 2,

$$M, \neg u_1 \wedge \neg u_2 \wedge \neg u_3 \wedge u_4 \models \neg a \wedge b$$

For the model after intervening to make $a = false$, we have:

$$M_{\neg a}, u_4 \models b \text{ and } M_{\neg a}, \neg u_4 \models \neg b$$

So when $do(\neg a)$ is true, b is true just when u_4 is true. Thus the SEM gives $P(b \mid do(\neg a)) = P(u_4) = p_4$, which justifies the use of the exogenous variable to represent the causal probability.

An exogenous variable is **extreme** if its probability distribution contains zeros, and is **non-extreme** if its probabilities are all strictly between 0 and 1.

In Pearl's semantics, the exogenous variables are assumed to be independent. However this semantics does not work even for simple cyclic models such as in Example 2 (when there are no interventions):

Proposition 1. *The noise variables $U_1, \ldots, U_4$ in the SEM of Example 2 cannot be non-extreme and independent.*

Proof. The instantiation $U_1 = true$, $U_2 = false$, $U_3 = false$, $U_4 = true$ is logically inconsistent, as it implies $(a \leftrightarrow b) \wedge (b \leftrightarrow \neg a)$, i.e,

$$M, u_1 \wedge \neg u_2 \wedge \neg u_3 \wedge u_4 \models false$$

and so $P(u_1 \wedge \neg u_2 \wedge \neg u_3 \wedge u_4)$ must have value 0. If the U_i are independent, this probability is

$$p_1(1-p_2)(1-p_3)p_4 \quad (3)$$

The only way for a product of real numbers to have value zero is for at least one of them to be zero. □

Note that there is another instantiation that also produces an inconsistency, namely $\neg u_1 \wedge u_2 \wedge u_3 \wedge \neg u_4$. The probability of this instantiation is:

$$(1-p_1)p_2p_3(1-p_4) \quad (4)$$

The non-existence of independent non-zero noise does not rely on there being two binary variables, but has to do with the cyclic causality. Suppose there is a discrete random variable X and an instantiation $\overline{u}$ to exogenous variables such that

$$M, \overline{u} \models X = v \rightarrow X = v'$$

where v' is a different value than v, (i.e, where one value implies another, perhaps through many causal steps) then

$$M, \overline{u} \models X \neq v$$

Thus we can derive the following:

Proposition 2. *If there is model M, a discrete variable X and an instantiation $\overline{u}$ to exogenous variables such that $P(u_i) > 0$ for each $u_i \in \overline{u}$ and for all values v of X,*

$$M, \overline{u} \models (X{=}v) \rightarrow (X{=}v')$$

where v' is different to v, then the exogenous variables cannot be probabilistically independent.

Note that proposition 2 does not hold for continuous variables, where it is possible that some single points have zero probability (for many distributions all individual points have zero probability). There has been much work on cyclic models with continuous variables that happily uses SEMs [Spirtes, 1995; Dash, 2005; Richardson, 1996; Lacerda *et al.*, 2008].

There seems to be three solutions to this problem of interpreting causal models with cycles:

- do not allow models that include cycles,
- make the noise dependent, or
- use a different semantics.

Example 4. One way to avoid inconsistency is to make the noise variables dependent, for example to make $u_2 \rightarrow u_1$, which makes $u_2 \wedge \neg u_1$ inconsistent. This can be modelled by making $u_2 = u_1 \wedge u_5$ for some noise u_5. Equation (1) becomes:

$$a \leftrightarrow (b \wedge u_1) \vee (\neg b \wedge u_1 \wedge u_5).$$

This can be reduced to $a \leftrightarrow (b \vee u_5) \wedge u_1$. This is the style of many of the SEMs of Pearl [2009], for example on page 29. This (with the corresponding equation for B) incorporates prior knowledge that A and B are positively correlated, as making one true can only increase the probability of the other being true. This is not appropriate if it is also possible that A and B are negatively correlated. It also does not result in a unique probability for A or for B.

Equilibrium Model Example

An alternative semantics is in terms of the equilibrium distribution (also called the stationary distribution) of a Markov chain [Brémaud, 1999]. For Example 1, this semantics is defined in terms of a Markov chain with variables $A^0, A^1, \ldots$ and $B^0, B^1, \ldots$, where the superscript represents a time point, with transition probabilities such as:

$$p_1 = P(a^t \mid b^t)$$
$$p_2 = P(a^t \mid \neg b^t)$$
$$p_3 = P(b^t \mid a^{t-1})$$
$$p_4 = P(b^t \mid \neg a^{t-1})$$

where a^t is the proposition that A is true at time t. This can be specified like an SEM, but variables on the right hand sides can refer to a previous time (in such a way that there are no cycles in the temporally extended graph). E.g.:

$$a^t \leftrightarrow (b^t \wedge u_1^t) \vee (\neg b^t \wedge u_2^t) \quad (5)$$
$$b^t \leftrightarrow (a^{t-1} \wedge u_3^t) \vee (\neg a^{t-1} \wedge u_4^t) \quad (6)$$

where for all t, U_i^t are independently identically distributed variables with probability p_i. The use of the previous time point avoids cycles in the temporally extended models. In the Markov chain, the A's at different times are different variables. There is no logical inconsistency that leads to the problem in the proof of Proposition 1.

The aim now is to determine the equilibrium distribution — the distribution over the variables that does not change in time. This Markov chain has an equilibrium that satisfies:

$$P(a) = p_1 P(b) + p_2(1 - P(b)) \quad (7)$$
$$P(b) = p_3 P(a) + p_4(1 - P(a)) \quad (8)$$

Solving the simultaneous equations gives:

$$P(a) = \frac{p_1 p_4 + p_2(1 - p_4)}{1 - (p_1 - p_2)(p_3 - p_4)} \quad (9)$$

$$P(b) = \frac{p_3 p_2 + p_4(1 - p_2)}{1 - (p_1 - p_2)(p_3 - p_4)} \quad (10)$$

which are well defined for all $p_i \in [0,1]$, except for the two cases: $p_1 = 1$, $p_2 = 0$, $p_3 = 1$, $p_4 = 0$ (which corresponds to $a \leftrightarrow b$) and $p_1 = 0$, $p_2 = 1$, $p_3 = 0$, $p_4 = 1$ (which corresponds to $a \leftrightarrow \neg b$). In these cases, there is an equilibrium for every value in $[0,1]$. For the rest of this discussion, we ignore extreme probabilities that give these two cases.

To specify Equations (5) and (6), we not only specified that A and B are dependent, but also that B depends on the *previous* value of A, and A depends on the *current* value of B. Intuitively, for each time, we sample B then A.

There are two main motivating reasons for adopting the equilibrium semantics:

- While we may not want to model time explicitly, there is often an underlying dynamical system where causes and effects happen. Because we have assumed non-extreme probabilities, all the Markov chains have an equilibrium distribution because they are ergodic and aperiodic. In the long run, samples from the dynamical system will be sampled from the equilibrium distribution.
- An equilibrium distribution is a belief state that will not be changed by more thinking. This is similar to the way equilibria are justified in game theory [Shoham and Leyton-Brown, 2008].

3 Markov Chain Equilibrium Models

In this section we define Markov chain equilibrium models as an alternative to SEMs for representing causal knowledge.

We assume finitely many discrete-valued variables. If X is a variable, the **parents** of X are defined to be a minimal set of variables $\mathbf{Y}$ such that for all sets of variables $\mathbf{Z}$ where $\{X\}$, $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint sets, the following condition holds:

$$P(X \mid do(\mathbf{Y})) = P(X \mid do(\mathbf{Y}, \mathbf{Z})).$$

That is, for all interventions where the variables in $\mathbf{Y}$ are set to particular values, changing the value of any other variables $\mathbf{Z}$ does not affect X. This is like the standard definition of conditional independence, but involves interventions, not observations. It is easy to show that the set of parents of X is unique.

We assume that all conditional probabilities are non-extreme. The non-extreme assumption is reasonable for learned models, where we may not want to a priori assume that any transition is impossible, but may not be appropriate for all domains. It simplifies the discussion as all of the Markov chains are then ergodic and aperiodic, with a unique equilibrium distribution, independent of the starting state [Brémaud, 1999].

This parent relation induces a directed graph that can contain cycles, but is irreflexive (there is no arc from a variable to itself).

Define a **causal network** to be an irreflexive directed graph where the nodes are random variables, together with a causal mechanism for each variable X that consists of a conditional probability $P(X \mid do(\pi_X))$ where π_X is the set of parents of X in the causal network.

To represent an intervention on a variable X, the causal mechanism for X is replaced by $P(X{=}v) = 1$ when we $do(X) = v$ [Pearl, 2009].

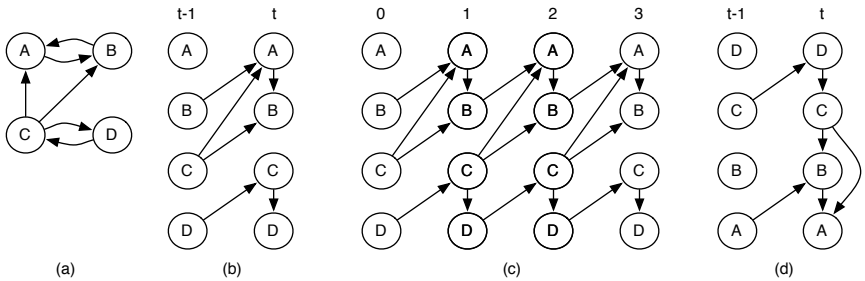

Figure 1: A causal network, its 2-stage DBN for sample ordering A,B,C,D, its unrolled DBN, and the 2-stage DBN for sample ordering D,B,C,D

To define the post-intervention semantics, we construct a two stage dynamic Bayesian network (DBN) [Dean and Kanazawa, 1989]. A 2-stage DBN specifies for each variable how the variable at the current stage depends on variables at the current stage and variables at the previous stage. This DBN depends on both the causal network and a **sample ordering** which is a total ordering of the variables. For each variable X, define π_X^- to be the set of those parents of X that are less than X in the sample ordering, and π_X^+ to be the set of those parents of X that are greater than X in the sample ordering. Thus $\pi_X = \pi_X^- \cup \pi_X^+$.

Intuitively, each X_i, depends on its parents in π_X^- at the current stage, and on its parents in π_X^+ at the previous stage.

A causal network with variables $\{X_1,\ldots,X_n\}$ and sample ordering $X_1,X_2,\ldots,X_n$ defines a decomposition of a discrete-time Markov chain where the state S^t at time t can be described by the variables $X_1^t,\ldots,X_n^t$ for each time t, and for each causal variable X for each time t, the Markov chain variable X^t has parents $\{Y^t : Y \in \pi_X^-\} \cup \{Y^{t-1} : Y \in \pi_X^+\}$. X^t is independent of all variables $Z^{t'}$ for $t' < t$, given these parents and is independent of all variables Z^t where $Z < X$ in the sample order given these parents in the Markov chain. Thus the causal network with the sample ordering defines the decomposition of the state transition function:

$$\begin{aligned} P(S^t \mid S^{t-1}) &= P(X_1^t,\ldots,X_n^t \mid S^{t-1}) \\ &= \prod_{i=1}^{n} P(X_i \mid X_1 \ldots X_{i-1} S^{t-1}) \\ &= \prod_{i=1}^{n} P(X_i \mid (\pi_{X_i}^-)^t (\pi_{X_i}^+)^{t-1}) \end{aligned} \tag{11}$$

The conditional probabilities for the Markov chain, the $P(X_i \mid (\pi_{X_i}^-)^t(\pi_{X_i}^+)^{t-1})$, are the $P(X \mid \pi_X)$ in the causal network.

Example 5. Consider the causal network in Figure 1 (a). In this example, the parents of A are B and C, the parents of B are A and C, the parent of C is D, and the parent of D is C.

Figure 1 (b) shows the 2-stage DBN with sample ordering A,B,C,D. The left nodes represent the variables at time $t-1$ and the right nodes represent the variables at time t.

The induced Markov chain is shown as a DBN in Figure 1 (c), where the structure is repeated indefinitely to the right. Each of the conditional probabilities is defined as part of the causal network. Note that in Markov-chain modeling this DBN is often given directly and is sufficient to model the distribution. This requires that the relationship between all the variables and their relative ordering are provided. The approach presented here deals with the problem of when the relations are provided, as with a cyclic causal model, but the ordering is not provided or is arbitrary.

Figure 1 (d) shows the 2-stage DBN for the same causal network with sample ordering D,C,B,A.

We define the distribution of the causal model (after interventions) to be the equilibrium (stationary) distribution of the induced Markov chain.

Given a causal network, when there are multiple sample orders under discussion, we will write the sample order as a subscript of the probability such as $P_{DCBA}(C)$.

4 Inference

The inference problem is: given a causal network and a sample ordering, determine $P(X|do(Y),Z)$ for some sets of variables X, Y and Z, which is the posterior distribution of X after doing Y and then observing Z in the equilibrium distribution[1]. This can be computed by replacing the causal mechanisms of the variables in Y with the intervention values, computing the equilibrium distribution, conditioning on Z and marginalizing over the remaining variables.

One way to compute the equilibrium distribution is to sample from it, sampling each variable in turn according to the sample ordering. This is an instance of Markov Chain Monte Carlo (MCMC) sampling [Brémaud, 1999]. In MCMC sampling we sample S^t from S^{t-1}, where S^t is the state at time t. The samples generated (after some burn-in period) can be considered as random samples from the equilibrium distribution, as long as there are sufficiently many.

MCMC can be carried out using Gibbs sampling. A state is an assignment of a value to each variable. If variables are selected according to the sample ordering, the probabilities from the causal model can be directly used to compute the equilibrium. To see this, suppose the sample ordering is $X_1,X_2,\ldots,X_n$, then we can use the decomposition of equation

[1]This is what Dash [2005] calls the manipulated-equilibriated model, but our equilibrium is over distributions. This is *not* counterfactual reasoning [Pearl, 2009], which would be observing then doing. In general, there could be arbitrary sequences of observing and doing, but that is beyond the scope of this paper.

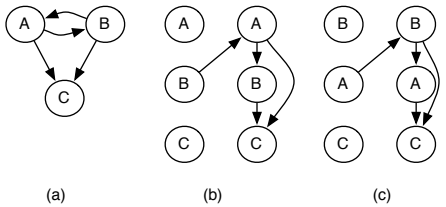

Figure 2: A causal network and two induced 2-stage Bayesian networks

(11), and note that, when X_i is selected, $(\pi^-_{X_i})^t(\pi^+_{X_i})^{t-1}$ are the current values of these variables. Thus Gibbs sampling of each variable using the probabilities of the causal network, sampled according to the sample ordering, produces samples from the equilibrium distribution.

Another approach to computing the equilibrium is to use an iterative inference method, starting with a probability distribution over states and repeatedly using the two-stage DBN to compute a distribution over the next state. Each variable can be updated using Equation (11). This converges to the stationary distribution (note MCMC just gives samples that are distributed according to the stationary distribution) with geometric convergence [Brémaud, 1999]. This algorithm is polynomial in state space (which is exponential in the number of variables) as it entails computing the probability of each state.

5 Dependence on Sampling Order

The following example shows that the equilibrium distribution can depend on the sample ordering:

Example 6. Consider the causal network of Figure 2 (a), with the causal probabilities:

$$P(a \mid do(b)) = 0.1 \;\; P(a \mid do(\neg b)) = 0.9$$
$$P(b \mid do(a)) = 0.9 \;\; P(b \mid do(\neg a)) = 0.1$$
$$P(c \mid do(a \wedge b)) = P(c \mid do(\neg a \wedge \neg b)) = 0.9$$
$$P(c \mid do(\neg a \wedge b)) = P(c \mid do(a \wedge \neg b)) = 0.1$$

Figure 2 (b) shows the 2-stage DBN with the sample ordering A, B, C. Figure 2 (c) shows the 2-stage DBN with the sample ordering B, A, C.

One way to think about this dependence on sample ordering is that doing B tends to change A to be different to B, yet doing A tends to change B to be the same as A. C has high probability if A and B have the same value. In the equilibrium distribution of Figure 2 (b), $P(c) = 0.82$, whereas in the equilibrium distribution of (c), $P(c) = 0.18$. Intuitively, in (b), A is sampled, then B is sampled, based on that value of A, and so they tend to have the same value and so C tends to be true. Whereas in (c), B is sampled, then A is sampled, based on that value of B, and so they tend to have different values and so C tends to be false.

By changing the probabilities in this particular causal network, for each sample ordering, the probability distribution of C in the equilibrium can be made to have an arbitrarily different non-extreme distribution.

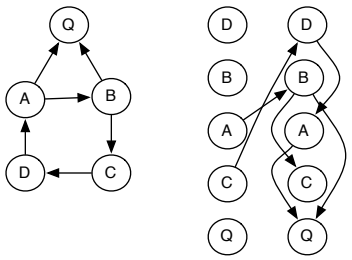

Figure 3: (a) a causal network and (b) a 2-stage DBN from Example 8

This was also observed by Lauritzen and Richardson [2002] in what they called feed-back models. In the rest of this paper we make some progress towards their challenge "It would be desirable to have a more precise understanding of the general relationship between the limiting distribution... and the conditional specification" [Lauritzen and Richardson, 2002, p. 346].

We can bound the dependence on the sample ordering as a function of the parameters. The causal network of Figure 2 (a) is a base prototypical case.

Proposition 3. *Given the structure of Figure 2 (a) and the causal probabilities of Example 1, the dependence of the probability of C on the sample ordering, namely* $|P_{ABC}(C) - P_{BAC}(C)|$ *is bounded as follows. For all values of C:*

$$|P_{ABC}(C) - P_{BAC}(C)|$$
$$\leq 2\left|\frac{p_1(1-p_2)(1-p_3)p_4 - (1-p_1)p_2p_3(1-p4)}{1-(p_1-p_2)(p_3-p_4)}\right|$$

Moreover this bound is tight; "$\leq$" approaches equality in the limit as C approaches the deterministic equivalence function, $C \leftrightarrow (A \leftrightarrow B)$ *or the exclusive-or,* $C \leftrightarrow (\neg A \leftrightarrow B)$.

See the appendix for a proof.

Note that the denominator is the same as the denominators of equations (9) and (10). It approaches zero (only) as the problem gets closer to the deterministic inconsistent model. The numerator is the difference between the probabilities (3) and (4). This is interesting because it is an exact characterization of the error, albeit a simple case.

The following two examples show what can occur with arbitrary orderings:

Example 7. Consider a directed model, where A is a coin toss, B represents whether someone who bet on heads cheers, and C is the hypothesis that A and B are related. Suppose the causal model is that A causes B and that together A and B cause C (with probabilities close to 1). Intuitively the correct sample order is A, B, C. The sample order B, A, C corresponds to a delay in the cheering; the cheering is for the previous coin toss. In this case, the current cheering (for the previous coin toss) is uncorrelated with the value of the current coin toss. If C were true when $A = B \wedge u_c$, as $P(u_c) \to 1$, the bound of Proposition 3 gives the exact difference in the probability of C between the sample orderings in the limit.

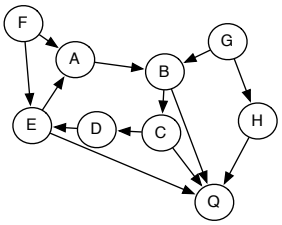

Figure 4: A causal network with a single cycle

Example 8. Consider the cyclic model of Figure 3(a) where A causes B, B causes C, C causes D, D causes A, and A and B together cause Q. In the sample ordering D, B, A, C, considering just these variables, the DBN has two disconnected components. The B and C variables at one time are connected to the A and D variables at the previous time, but not at the current time (see Figure 3(b)). Thus A and B are unconditionally independent (given no observations) in the equilibrium.

We can define a restricted class of sample orderings that is better behaved than arbitrary orderings. Given a causal network, a **faithful sample ordering** is a total ordering of the variables such that

- for every chain in the causal network that is not part of a cycle, the parents of a variable occur before the variable in the total order, and
- for every cycle in the causal network, for all but one of the variables in that cycle, the parents of a node occur before that node in the total ordering.

Consider the case where, after all interventions, there is a single directed cycle between a set of variables, and a variable Q that depends (perhaps indirectly) on some subset of the variables in the cycle. There can be other parents of the variables in the cycle and other parents of Q, but no other cycles. An example is shown in Figure 4.

Proposition 4. *Suppose there is a causal network with a single directed cycle, and σ_1 and σ_2 are two faithful sample orderings where X_j is the variable in the cycle that is before it parents in σ_1 and X_k is the variable that is before its parents in σ_2, then for any proposition Q,*

$$|P_{\sigma_1}(Q) - P_{\sigma_2}(Q)| \leq 2|P(X_j)P(X_k \mid do(\pi_{X_k})) - P(X_k)P(X_j \mid do(\pi_{X_j}))|$$

See the appendix for a proof.

One thing to notice is that if the variables in the cycle are symmetric in that each variable in the cycle depends on its parents in the same way, then the difference between the faithful sample orderings is zero.

When there are multiple interacting cycles, the situation can become very complicated. We investigate one such situation below.

6 Evaluation

In order to determine the effect of sample ordering on more complicated domains, we investigated a domain which is an abstraction of a larger class of domains and yet small enough to be able to compute the equilibrium exactly (to within machine precision) by iterating through the transition dynamics until convergence.

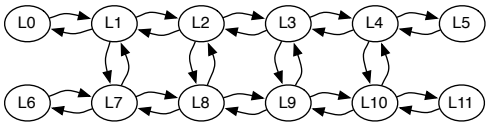

Figure 5: Network N : A spatial example

Example 9. We investigate the sensitivity to sample ordering for a spatial domain with 12 locations shown in Figure 5. There is a Boolean random variable at each location which depends on the variables at its neighbouring locations. We assume a form of exchangeability where each node with the same connectivity has the same probability dependence on its neighbours and the neighbours are exchangeable. Thus, each location depends causally on its neighbours in the same way. The probability of the variable at each location depends on the number of its neighbours that are set to true. For the locations with three neighbours, the causal distribution is defined by 4 parameters, $p_0, \ldots, p_3$ where p_i is the probability that the variable at the location is true given that i of its neighbours are set to true and the others are set to false. For the locations with one neighbour, its causal probability is governed by two numbers; p_t, the probability given is its neighbour is set to true, and p_f, the probability given its neighbour is set to false. Thus this model has six real-valued parameters. These are reasonable assumptions for a spatial domain, and an acyclic sample ordering (or even a probabilistic mix of acyclic models, which may be required to preserve symmetry) does not adequately reflect the domain.

This is a natural domain to investigate dependence on sample ordering, because symmetry considerations mean that some probabilities should be identical. However, a sample ordering can break the symmetry. Thus by comparing the probabilities of nodes that are symmetric, we can investigate the dependence on sample ordering.

To compute the equilibrium, starting from a random state, we used the iterative method described above. This involved repeatedly adding a new current stage and summing out the old previous stage, until the probabilities changed by less than 10^{-15} and then ran 20% more steps. This entailed representing a distribution over the 2^{12} states. We then considered any difference of less than 10^{-15} to be zero (assuming it to be a rounding error). Note that the results here are many orders of magnitude more accurate than could be detected by MCMC in any reasonable time.

Empirically, over hundreds of runs, in 29.2% of the orderings[2], $L1$, $L4$, $L7$ and $L10$ all have the same marginal probability. The actual parametrization did not affect whether these were equal. Let's call the class of ordering where these 4 variables have the same marginal Class 0, and the class containing the other orderings Class 1.

For those not in class 0, in 20% of the orderings $L1$ and

[2]All statistics in this section are plus or minus 2% with 95% confidence.

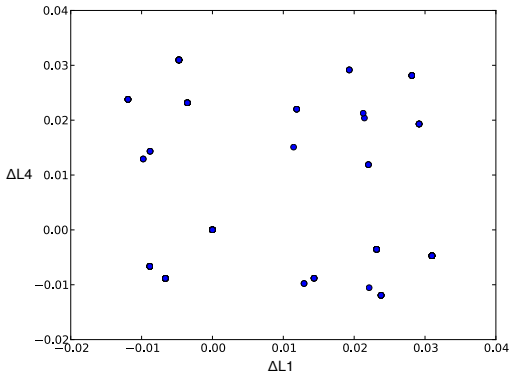

Figure 6: Varying sample order for a fixed parameterization. See text for a description of the axes.

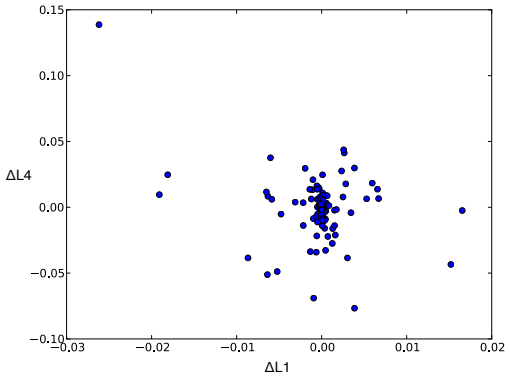

Figure 7: Varying parameterizations for fixed sample ordering

$L10$ have the same probability that is different from $L4$ and $L7$. In 6% of all the orderings, $L1$ and $L4$ have the same probability even when the rest of the variables are not equal to each other. Whether the symmetries arose for a particular ordering was independent of the parametrization. In none of the orderings where the locations $L1$, $L4$, $L7$ and $L10$ were not all the same did $L1$ and $L7$ have the same probability.

To understand the effect of sample ordering, we first investigated the probability of $L1$ in relationship to $L4$, which, by symmetry should have the same probability, and indeed have the same probability in approximately 35% of the cases. Figure 6 shows the variability of sample order for a representative fixed parametrization[3] and 200 sample orderings. First, we ran the sample order $L0, L1, \ldots, L11$, to be the reference order. For each random order, on the x-axis we plot $\Delta L1$, the value of $L1$ for the random order minus the value of $L1$ in the reference order, and on the y-axis we plot $\Delta L4$, the value of $L4$ for the random order minus the value of $L4$ in the reference order. As $L1$ and $L4$ have the same value in the reference order, this lets us see the variability of predictions, and see when $L1$ and $L4$ gave different predictions (when they are off diagonal). Note that there are many cases at $(0,0)$, but there is essentially no structure for the cases where the ordering does not imply they are identical.

Figure 7 shows a plot with the same meaning, but with a fixed sample ordering, (in this case $[L9, L1, L11, L2, L0, L10, L7, L4, L8, L5, L3, L6]$) and 200 random parametrizations. For this ordering, no pair of locations $L1$, $L4$, $L7$ and $L10$ had identical probabilities. Again there is no apparent structure in the relationship between $L1$ and $L4$.

Next we approached the question of what properties of the ordering characterize class 0, where the symmetry is preserved. The network N contains three square arrangements of nodes, namely $L1, L2, L7, L8$; $L2, L3, L8, L9$; and $L3, L4, L9, L10$. Each ordering of the variables provides a direction to the arcs between the nodes in each of the squares, where each variable has arcs from its neighbours that are predecessors in the ordering; this provides the belief network of the "current" nodes. There are three unique acyclic patterns that can be induced on a four node cycle, these are shown in Figure 8.

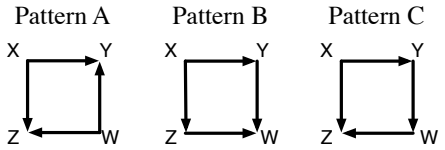

Figure 8: All orderings result in just three unique patterns within each square for network N, up to rotation and mirror images. There are 8 orderings for each pattern; e.g., for pattern A, there are four choices for X, then comes its opposite node W, then there are two choices as to whether Y or Z comes next. Similarly for pattern B, there are 4 choices for X then two choices for its neighbour, then the other neighbour and the node opposite to X follow. For pattern C there are 4 choices as to which node is the sink, and also a choice of direction (clockwise or counterclockwise).

It turns out empirically that an ordering is in Class 0 if and only if all of the square patterns induced by the ordering are of pattern A or pattern B. All orderings that are in Class 1 have at least one instance of pattern C.

When choosing an ordering for the entire network (excluding the end nodes) each square arrangement of nodes will have one of these patterns induced upon it. Note that only orderings which match directions on the overlapping arcs for $(L8, L2)$ and $(L9, L3)$ are possible. Counting these up for all assignments, there are 16 ways to choose patterns A or B for the left square and 8 for each of the other two squares, as they need to match with $(L8, L2)$ or $(L9, L3)$, so there are $(16 \times 8 \times 8) = 1024$ combinations using only A or B. To count the total number of consistent assignments we can choose the first cycle any of 24 ways, then each subsequent cycle has 12 remaining patterns which are consistent, yielding $24 \times 12 \times 12 = 3456$ consistent orderings. The expected proportion of a set of random orderings that would only use patterns A and B is thus $1024/3456 \approx 29.6\%$ which is very close to what we see empirically for Class 0.

[3] $p_0 = 0.06680, p_1 = 0.64617, p_2 = 0.86093, p_3 = 0.09159, p_f = 0.95610, p_t = 0.41803$

7 Conclusion

If SEMs are the right model for causality in cyclic domains, they should work for simple cases. In this paper, we have argued that they impose undesirable dependencies, and proposed MC equilibrium models as an alternative.

The idea that a causal model means the equilibrium of some sort is not new; Strotz [1960] argued that when there are variables that are interdependent in a cyclic ordering, the fixed point in values was a specification error. Others (e.g., Fisher [1970]) followed up by giving conditions for the equilibrium to be well defined. Spirtes [1995] gives the equilibrium interpretation but uses SEMs with continuous variables. There is also a literature on learning cyclic causal models that focusses on the interpretation that causal cycles are caused by unmeasured latent variables [Glymour and Spirtes, 1988; Schmidt and Murphy, 2009]. Lauritzen and Richardson [2002] give a Markov chain equilibrium semantics for the feed-back models for chain graphs. They leave the relationship between the ordering and the resulting distribution as an open problem, which we have made progress on, both theoretically in Propositions 3 and 4 and empirically for a class of parametrized models.

It should also be noted that the counterexample of Neal [2000] to Pearl and Dechter [1996] does not work for the equilibrium semantics. D-separation holds with the MC equilibrium semantics as it uses a directed network.

Users of causal models where cycles over time are possible should be aware that the causal ordering is important information of a different kind than the local causal relations between variables. In some domains the ordering may seem obvious in which case a full DBN may be the appropriate representation. However, if the ordering is in fact arbitrary or unknown, as with spatial policies or in relational domains, the choice of ordering can have a large impact on the resulting marginal distributions. Sample ordering can provide extra flexibility in fitting interventional data and observational data. Note that while this paper has assumed a fixed sample ordering, a distribution over sample orderings will allow for more flexible modelling (e.g., in Example 6, mixtures of orderings can give any probability in the range $[0.18, 0.82]$ rather than just at the end points of this range).

Appendix

Proof of Proposition 3

Pick a particular value of c of C. In the ABC ordering, the DBN uses $P(B \mid do(A)$ for B and the equilibrium distribution for A, and analogously for the other ordering. Thus,

$$\begin{aligned}
&P_{ABC}(c) - P_{BAC}(c) \\
&\quad = \sum_{AB} P(c \mid AB)((P_{ABC}(AB) - P_{BAC}(AB)) \\
&\quad = \sum_{AB} P(c \mid AB)(P(A)P(B \mid do(A)) - P(B)P(A \mid do(B)))
\end{aligned}$$

Consider the $A = true, B = true$ case. Using equations (9) and (10) for $P(A)$ and $P(B)$, let

$$\begin{aligned}
\alpha &= P(a)P(b \mid do(a)) - P(b)P(a \mid do(b)) \\
&= \frac{p_1 p_4 + p_2(1-p_4)}{1-(p_1-p_2)(p_3-p_4)} p_3 - \frac{p_3 p_2 + p_4(1-p_2)}{1-(p_1-p_2)(p_3-p_4)} p_1 \\
&= \frac{p_1(1-p_2)(1-p_3)p_4 - (1-p_1)p_2 p_3(1-p4)}{1-(p_1-p_2)(p_3-p_4)}
\end{aligned}$$

Similar calculations for the other values gives:

$$\begin{aligned}
&P_{ABC}(c) - P_{BAC}(c) \\
&\quad = P(c \mid ab)\alpha + P(c \mid a\neg b)(-\alpha) \\
&\qquad + P(c \mid \neg b)(-\alpha) + P(c \mid \neg a \neg b)\alpha
\end{aligned}$$

Consider the parameters of $P(c \mid AB)$ as free. This is a linear equation of these free parameters, and so has maxima and minima at the extreme values of these parameters (which is 0 and 1). A simple enumeration of these gives a maximum value of $|P_{ABC}(c) - P_{BAC}(c)|$ when $P(c \mid ab) = P(c \mid \neg a \neg b) = 1$ and $P(c \mid \neg ab) = P(c \mid a \neg b) = 0$. The maximum is $2|\alpha|$.

Proof of Proposition 4

If Q is a complex proposition on more than one variable, we can construct a new variable that is true whenever Q is true.

If Q does not have any variables in the cycle as ancestors, the proposition trivially holds. Thus the only remaining case is when Q is represented as a variable where an element of the cycle is an ancestor.

Suppose the ancestors of Q are $X_1 \ldots X_m$ (which includes all of the variables in the cycle). The only conditional probabilities that have different values in the induced DBNs under σ_1 and σ_2 are the probabilities defining X_j and X_k. The DBN induced by σ_1 uses the value from the equilibrium for X_j and uses the causal conditional probability for X_k, and the DBN induced by σ_2 uses the causal conditional for X_j and the equilibrium distribution for X_k.

$$\begin{aligned}
&P_{\sigma_1}(Q) - P_{\sigma_2}(Q) \\
&\quad = \sum_{X_1 \ldots X_m} P_{\sigma_1}(Q \mid X_1 \ldots X_m) P_{\sigma_1}(X_1 \ldots X_m) \\
&\qquad - \sum_{X_1 \ldots X_m} P_{\sigma_2}(Q \mid X_1 \ldots X_m)_{\sigma_2} P(X_1 \ldots X_m) \\
&\quad = \sum_{X_1 \ldots X_m} \begin{array}{l} P(Q \mid \pi_Q)(\prod_{i \notin \{j,k\}} P(X_i \mid \pi_{X_i})) \\ (P(X_j)P(X_k \mid do(\pi_{X_k})) - P(X_j \mid do(\pi_{X_j}))P(X_k)) \end{array}
\end{aligned}$$

because all of the common factors can be distributed out of the difference. Summing out the variables in $X_1 \ldots X_m$ other than X_j and X_k gives

$$\begin{aligned}
&P_{\sigma_1}(Q) - P_{\sigma_2}(Q) \\
&\quad = \sum_{X_i, X_k} \begin{array}{l} P(Q \mid X_j X_k) \\ (P(X_j)P(X_k \mid do(\pi_{X_k})) - P(X_j \mid do(\pi_{X_j}))P(X_k)) \end{array}
\end{aligned}$$

which is the same case as in the proof of Proposition 3. Note that summing out variables does not always result in a conditional probability, but it does in this case. As in the proof of Proposition 3, treating $(P(X_j)P(X_k \mid do(\pi_{X_k})) - P(X_j \mid do(\pi_{X_j}))P(X_k))$ as fixed, this is a linear equation of the parameters for Q, and so is bounded at the extreme values. Enumerating the cases gives the bound in the proposition.

References

[Arnold *et al.*, 2001] Barry C. Arnold, Enrique Castillo, and José María Sarabia. Conditionally specified distributions: An introduction. *Statistical Science*, 16(3):249–274, 2001.

[Brémaud, 1999] Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*. Springer, 1999.

[Crowley and Poole, 2011] Mark Crowley and David Poole. Policy gradient planning for environmental decision making with existing simulators. In *Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI-11), special track on Computational Sustainability and AI*, San Francisco, 2011.

[Dash, 2005] Denver Dash. Restructuring dynamic causal systems in equilibrium. In *Proc. 10th Int. Workshop on AI and Stats*, pages 81–88, 2005.

[Dean and Kanazawa, 1989] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150, 1989.

[Domingos *et al.*, 2008] Pedro Domingos, Stanley Kok, Daniel Lowd, Hoifung Poon, Matthew Richardson, and Parag Singla. Markov logic. In L. De Raedt, P. Frasconi, K. Kersting, and S. Muggleton, editors, *Probabilistic Inductive Logic Programming*, pages 92–117. Springer, New York, 2008.

[Fisher, 1970] Franklin M. Fisher. A Correspondence Principle for Simultaneous Equation Models. *Econometrica*, 38(1), 1970.

[Gelman, 2004] A. Gelman. Parameterization and bayesian modeling. *J. Am. Stat. Assoc.*, 99(466):537–545, 2004.

[Glymour and Spirtes, 1988] C. Glymour and P. Spirtes. Latent variables, causal models and overidentifying constraints. *Journal of Econometrics*, 39(1-2):175–198, 1988.

[Halpern, 2000] Joseph Y. Halpern. Axiomatizing causal reasoning. *Journal of AI Research*, 12:317–337, 2000.

[Heckerman *et al.*, 2000] D. Heckerman, D.M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *J. Mach. Learn. Res.*, 1:49–75, 2000.

[Iwasaki and Simon, 1994] Yumi Iwasaki and Herbert A. Simon. Causality and model abstraction. *Artificial Intelligence*, 67(1):143–194, 1994.

[Lacerda *et al.*, 2008] Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 366–374. AUAI Press, 2008.

[Lauritzen and Richardson, 2002] S.L. Lauritzen and T.S. Richardson. Chain graph models and their causal interpretations. *J. R. Stat. Soc. B*, 64(3):321–361, 2002.

[Neal, 2000] R. M. Neal. On deducing conditional independence from d-separation in causal graphs with feedback (research note). *JAIR*, 12:87–91, 2000.

[Neville and Jensen, 2007] Jennifer Neville and David Jensen. Relational dependency networks. *Journal of Machine Learning Research (JMLR)*, 8:653–692, 2007.

[Pearl and Dechter, 1996] Judea Pearl and Rina Dechter. Identifying independencies in causal graphs with feedback. In *Proc. 12th Conference on Uncertainty in AI*, pages 420–426, 1996.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[Richardson, 1996] Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Proceedings of the Twelfth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pages 454–461, San Francisco, CA, 1996. Morgan Kaufmann.

[Schmidt and Murphy, 2009] Mark Schmidt and Kevin Murphy. Modeling discrete interventional data using directed cyclic graphical models. In *Proc. 25th Conf. on Uncertainty in AI*, pages 487–495, 2009.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game Theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[Spirtes, 1995] Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 491–498, San Francisco, CA, 1995. Morgan Kaufmann.

[Strotz and Wold, 1960] Robert H. Strotz and H. O. A. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica*, 28(2):417–427, 1960.

[Strotz, 1960] Robert H. Strotz. Interdependence as a specification error. *Econometrica*, 28(2):428–442, 1960.