A Note on Cyclic Graphs and Dynamical Feedback Systems

Thomas Richardson
(University of Washington; tsr@stat.washington.edu)
Peter Spirtes
(Carnegie Mellon University; ps7z@andrew.cmu.edu)
Clark Glymour
(University of California, San Diego and Carnegie Mellon University;
cg09@andrew.cmu.edu)

1. Introduction. Directed acyclic graphical (DAG) models were motivated in large part by the desire to have a general formalism to represent causal hypotheses and the restrictions on probability distributions they imply. DAG models exploited a fundamental kinship in a variety of statistical formalisms often treated as distinct "models": factor models, structural equation models, regression models, logistic regression models, survival models, etc. The fundamental connection is through either of two equivalent (for DAGs) properties, a "local" Markov condition, or the property of d-separation, sometimes called the "global" Markov condition. (Pearl 1988, Lauritzen *et al.* 1990). In much the same spirit, directed cyclic graphs (DCGs) have been introduced to represent the causal structure of feedback processes and the restrictions on probability distributions those structures imply. Developments in our understanding of DCGs have proceeded so rapidly that it is appropriate to consider the prospects and limitations of cyclic representations of feedback systems. (For an alternative approach to extending graphical models to the temporal domain see Aliferis and Cooper, 1996)

2. DCG Properties. The causal claims of recursive linear "structural equation" models, and the statistical properties they imply, are nicely captured by DAGs, and in both economics and engineering certain feedback systems are modeled by linear equations. (For a discussion of the relation between DAGs and linear structural equation models, see Spirtes et. al 1993). It is natural therefore to first consider cyclic graphs for simultaneous linear equations with independent errors. The local Markov condition--variables are independent of their non-descendants conditional on their parents--fails. Spirtes (1994) and Koster (1996) showed that d-separation entails conditional independence, and Spirtes(1994) showed that if a DCG entails a conditional independence then the corresponding d-separation relation holds in the linear DCG. Richardson (1996a) gave a polynomial time decision procedure for equivalence and a correct discovery procedure from a conditional independence oracle (Richardson 1996b). While issues about latent variables and selection bias remain, in principle our understanding of the linear cyclic case is now almost as good as our understanding of the linear acyclic case.

For DCGs, non-linear feedback systems present a more complicated picture. Spirtes (1994) showed by example that d-separation does not entail conditional independence for non-linear systems with continuous variables, but showed that a weaker condition (in which cycles are made into cliques and d-separation applied to the resulting acyclic graph) gives (not necessarily all) conditional independence relations implied by the graph and the functional form. Pearl and Dechter (1996) showed that in models with discrete variables, in which each vector of values of the errors determines a unique value of the other variables, i.e. which have a reduced form, and with independent errors, d-separation in the corresponding cyclic graph implies conditional independence. The converse is not true in general.

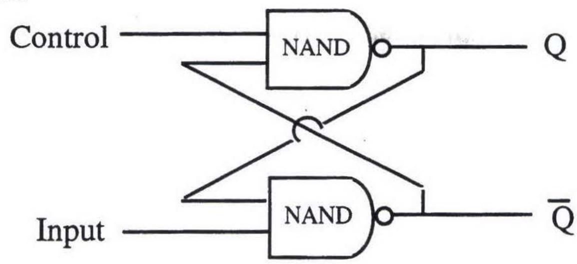
The incompleteness of d-separation for non-linear DCGs, whether discrete or continuous, is a challenge to the possibility of learning much about such structures from data. Available means for learning information about DAGs from data, no matter whether by Bayesian or conditional independence techniques, depend on the completeness of d-separation and the still stronger assumption that special parameter values do not create conditional independence. In the cyclic case we need, at the very least, an understanding of what if any partial information about graphical structure can be extracted from the probabilities when d-separation is incomplete.

3. Causal Interpretations of DCGs. DCGs naturally suggest a stochastic process, or time series, but the correspondence is unclear. A tradition in econometrics has investigated the connection between"non-recursive" linear simultaneous equation models--which we now know can be represented by DCGs--with time series whose causal claims can be represented by infinite acyclic graphs. The conditional independencies implied by any linear DCG are also implied by a time series in which each variable X_t receives the same, constant, exogenous shock at all times t. It has been claimed that these are the *only* time series preserving the global Markov properties of DCGs, but we know of no proof. If so, linear DCGs, and therefore also linear non-recursive structural equation models, form a rather restricted class of models.

4. Relaxing the Assumption of the Existence of a Reduced Form. Many naturally occurring dynamic systems and digital circuits do not possess a reduced form. This can happen in one of two ways: there may be no solution for the measured variables compatible with a given setting of the error variables, or there may be more than one solution. For example, consider a model with structure 1 below in which each variable is binary with the following functional relationships (using binary addition):
$B := C + A + \varepsilon_B$; $C := B + D + \varepsilon_C$; $A := \varepsilon_A$; $D := \varepsilon_D$. If the error terms are all 0, ($\varepsilon_A = \varepsilon_D = \varepsilon_B = \varepsilon_C = 0$) then there are multiple solutions for B and C: B=C=1, and B=C=0.

Pearl and Dechter find a rather complicated discrete parametrization of graph (1) that is determinate, and we have found others, but in our (brief) experience, most discretizations with most simple functional forms on cyclic graphs result in systems with indeterminate states. Again, no natural measure of the proportion of such cases is known.

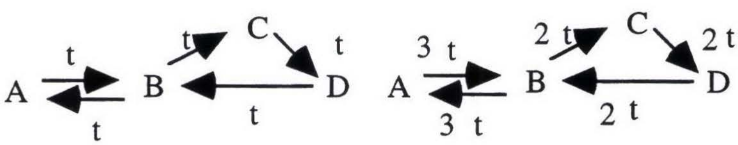
Pearl and Dechter suggest that indeterminate discrete systems are "unstable"--small variations in the values of exogenous or error variables produce big variations in other variables. It is not clear to us what that means in the case of categorical variables, and in any case the same "instability" arises in non-linear acyclic models. Moreover, feedback structures with multiple equilibria seem to occur in applied sciences. For example, in the bistable memory circuit:



if Input and Control are both 1, then there are two equilibria for Q and $\bar{Q}$:Q=1, $\bar{Q}$ =0, or Q=0, $\bar{Q}$ =1. Contrary to the econometric literature on non-linear systems, which restricts itself to the case in which there are unique equilibria (Goldfeldt & Quandt (1972)), we suggest that such systems should be studied in more detail: systems with multiple equilibria or no equilibrium do appear to be a fundamental feature of discrete, cyclic models.

5. Rates and mechanisms. Directed graphs with cycles, interpreted causally, invite us to think of equilibrium instantiations of a feedback system arising from an initial state by a temporal sequences of causal steps described by the graph and the functional dependencies. Under such an interpretation the equilibria that the model arrives at is independent of the initial state of the (endogenous) non-error variables, since it is uniquely by the (exogenous) error terms. However, in examples such as the bistable circuit above the equilibria also depend upon the initial state of the endogenous variables.

Moreover, in more complex systems with time delays or varying rates of propagation, more than the initial state at one moment may be necessary in order to determine a sequence of successive states. And in many real cases, rates of propagation, determine the outcome. Consider those intricate rows of dominoes whose direction of fall depends on races between other rows of falling dominoes. In complex systems, e.g. in engineering or medicine, the influence of one variable on another may take some time to propagate, and different influences may have different rates. Furthermore in feedback systems such rates can be critical not only to the equilibrium distribution but to whether any limiting state is obtained at all. Consider for example two models it which the attachments to edges indicate the lag in units t:



i ii

where in case (i) $A_t := B_{t-1}$; $C_t:=B_{t-1}$; $D_t: = C_{t-1}$; $B_t := A_{t-1} \bullet D_{t-1}$, and in case (ii) $A_t := B_{t-3}$; $C_t:=B_{t-2}$; $D_t:=C_{t-2}$; $B_t:= A_{t-3} \bullet D_{t-2}$ (binary multiplication).

If system (i) is initialized with at least one variable with value 0 then every variable eventually becomes 0; otherwise everything stays at 1. If we imagine a population of such systems with any distribution of start states, after finite time the values of all variables are correlated, and any two variables are independent conditional on any third variable. In case (ii) the state at three successive time units may be required to specify a unique sequence of states, for example

Initial States

Times	**A**	**B**	**C**	**D**
t	1	0	0	0
t+1	0	0	0	1
t+2	0	0	0	0

In this case the sytstem (ii) never converges but oscillates, each variable taking the value 1, for one time unit out of every six. This suggests a need to understand the statistical properties of operator graphs--directed graphs with time operators attached to edges indicating the lag of the dependency. Such graphs are considered by Heise (1975) and in the linear systems and control literature in engineering (see for example Huggins & Entwistle 1968). A dynamic operator graph (DOG) is a finite representation of an infinite stochastic time series process. Associated with each edge in the graph is a polynomial in the operator 't'. If we associate a paremeter of 'αt^n' with a given edge A→B, then this means that, if all other variables were held constant for the duration, then an increase in the value of A by one unit at time t_0, would result in an increase of size α in the value of B, at time t_0+n. Thus the time series equation:

$$Y_t = \alpha \cdot X_{t-1} + \varepsilon_{Y_t} \text{ is represented as } X \xrightarrow{\alpha} Y$$

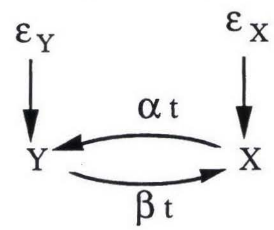
We make the following convention; If a variable X in an operator graph is without parents, then the sequence $(X_t\}$ is an i.i.d. sequence of normally distributed random variables.

When combined with a probability distribution over the intial state sets, an operator graph specifies, among other things, a time series. The time series may or may not converge to a limiting joint probability distribution on the variables, depending on the graph, the lag structure, the functional form and parameters, the initial probability distribution, and how the limit is taken. The details of these connections do not seem to be well understood. However, we will present a preliminary result related to calculating the equilibrium distribution of cyclic linear operator graphs subject to independent random shocks.

Corresponding to a time series, there is an equivalent acyclic, infinite graphical representation of any given dynamic operator graph, in which the variables are indexed by time. This infinite graph can be constructed from the operator graph, in the following way:

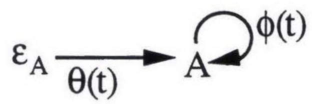
a) For each variable X in the DOG, assign an infinite sequence of variables $\{X_t\}$, $t \in \mathbf{Z}$.

b) For each edge from X to Y, X→Y, in the DOG, with coefficient p(t), do the following: If the coefficient of t^n in p(t) is non-zero, then in the corresponding infinite graph place an edge between X_t and Y_{t+n}, for every t. The following DOG:



can thus be represented by the following infinite DAG:

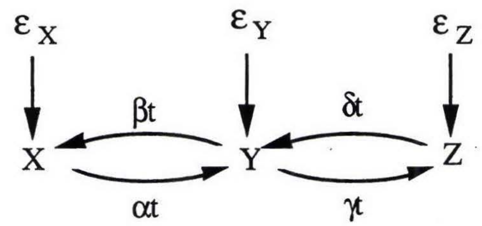
We must also allow for 'self-loops'; if A is a cause of B, and B is a cause of A, but we fail to measure B, then A (in one time period) is a cause of A (in a later time period). This is represented in a DOG as an edge which begins and ends at a given vertex. Using the DOG notation it is possible to represent a large class of models, for example *any* univariate auto-regressive moving average (ARMA) model, as follows:



However, intuitions concerning causal systems with feedback suggest restricting our attention to subclasses in which various restrictions hold. In most, if not all, cases in which we speak of (a change in) A causing (a change in) B the cause preceeds the effect. This can easily be represented in the DOG notation by requiring that in every edge polynomial p(t) the coefficient of t^0 is 0, i.e. the polynomial has no constant term. We might wish to drop this condition if our main concern is with approximating a certain causal mechanism. For example, some causal influences may be propagated several orders of magnitude faster than others, in which case, particularly if practical considerations force us to take measurements at widely spaced intervals, we may wish to construct a model in which a change in one variable immediately produces a change in another.

Another restriction is suggested by a strange feature of the example above of a bivariate DOG: it can in fact be decomposed into two disjoint systems, the first consisting of X_{2t}, and Y_{2t+1}, the second consisting of Y_{2t}, and X_{2t+1}. One very weak constraint that can be imposed to avoid such an occurence is the assumption of a Weak Trek Connection: If X_t is correlated with Y_s and $Y_{s'}$, (t^2 s,s') then X_t is correlated with Y_r, where s ^{2} r ^{2} s'. i.e. if X at time t is correlated with Y at times s and s', then X at time t is also correlated with Y during the interval in between. A slightly stronger condition would require Strong Trek Connection: If X_t is correlated with Y_s (t^2s) then X_t is correlated with Y_r, for r < s, i.e. if X at time t is correlated with Y at time s, then X at time t is also correlated with Y at all prior time points. This list or regularity conditions is of course not supposed to be exhaustive.

6.Mason's Rule applied to Dynamic Operator Graphs. Mason's rule--the standard technique for calculating effects and correlations in linear non-recursive simultaneous equation systems (see Heise, 1975)--can be applied to obtain a reduced form for a given variable in terms of the exogenous error terms. If we apply Mason's rule to an operator graph we can express each measured variable as a finite sum of power series in t, one power series for each error term. Consider the following DOG:



By way of example, if we were to apply Mason's Rule to the model, to calculate the gain from ε_X, ε_Y and ε_Z to X, we arrive at the following expression:

$$X = \frac{\varepsilon_X + \beta t\, \varepsilon_Y + \delta\beta t^2\, \varepsilon_Z}{1-(\alpha\beta+\gamma\delta)t^2}$$

which can then be expressed as a power series, assuming that $|\alpha\beta+\gamma\delta| < 1$:

$$X = \sum_{s=0}^{\infty}(\alpha\beta+\gamma\delta)^s t^{2s}[\varepsilon_X + \beta t\cdot\varepsilon_Y + \delta\beta t^2\cdot\varepsilon_Z]$$

Now, considering t as an operator, such that t: X_s ♋ X_{s-1}, and t^r: X_s ♋ X_{s-r}, we can derive the following expression for X_τ in terms of $\varepsilon_{X_s}, \varepsilon_{Y_s}, \varepsilon_{Z_s}$ $(s < \tau)$

$$X_\tau = \sum_{s=0}^{\infty}(\alpha\beta+\gamma\delta)^s[\varepsilon_{X\tau-2s} + \beta\cdot\varepsilon_{Y\tau-2s-1} + \delta\beta\cdot\varepsilon_{Z\tau-2s-2}]$$

Hence we can express X_τ, a variable in the extended time series corresponding to the DOG in terms of the exogenous error terms in that time series. This is equivalent to solving for X_τ in terms of $\varepsilon_{X_s}, \varepsilon_{Y_s}, \varepsilon_{Z_s}$ $(s < \tau)$ from the following three recursive equations:

$$X_\tau = \beta \cdot Y_{\tau-1} + \varepsilon_{X\tau}$$
$$Y_\tau = \alpha \cdot X_{\tau-1} + \delta \cdot Z_{\tau-1} + \varepsilon_{Y\tau}$$
$$Z_\tau = \gamma \cdot Y_{\tau-1} + \varepsilon_{Z\tau}$$

Note that all we have done in order to go from the reduced form in the DOG to the reduced form in the corresponding time series model is effectively to write the exponent of the 't'-coefficient for a given variable occurring in the first expression as the lag from τ by which the corresponding variable is indexed in the second expression.

We can, in this way, express every variable as a power series in terms of the prior error terms. Given the assumption that the error terms are independent (both for different times, and different variables), and i.i.d. it is then possible, in principle, to calculate the limiting covariance between any two variables. For example suppose that:

$$X_t = \sum_{V \in V} \sum_{s=0}^{\infty} \alpha_{Vt-s} \varepsilon_{Vt-s}, \text{ and } Y_t = \sum_{V \in V} \sum_{s=0}^{\infty} \beta_{Vt-s} \varepsilon_{Vt-s}$$

under the assumption on the error terms, it then follows that:

$$\mathrm{Cov}(X_t, Y_t) = \sum_{V \in V} \sum_{s=0}^{\infty} \alpha_{Vt-s} \beta_{Vt-s} V(\varepsilon_V)$$

Thus we can derive the limiting covariance matrix from the DOG, via Mason's rule.

References

ALIFERIS, C., G.F. COOPER (1996). A structurally and temporally extended Bayesian belief network model: Definitions, Properties and Modeling Techniques In *Uncertainty in Artificial Intelligence, Proceedings, 12th Conference*, Morgan Kaufman, CA

GOLDFELD, S.M., R.E.QUANDT (1972). Non-linear Methods in Econometrics. Contributions to Economic Analysis 77. North-Holland, Amsterdam.

HEISE D.(1975). Causal Analysis. Wiley, New York.

HUGGINS, W.H. and ENTWISTLE, D.R.(1968) Introductory Systems and Design. Ginn Blaisdell, Waltham Massachussetts.

KOSTER, J.T.A.(1996) Markov Properties of Non-Recursive Causal Models. Ann. Statist **24**

LAURITZEN, S., A.P. DAWID, B. LARSEN, H. LEIMER, (1990). Independence properties of directed Markov fields, *Networks* **20**

LAURITZEN S.L., D.J. SPIEGELHALTER, (1988). Local computation with probabilities in graphical structures and their applications to expert systems (with discussion). *JRSS B*, **50**, No.2.

PEARL, J. (1988). Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, CA.

PEARL, J., R. DECHTER (1996). Identifying independencies in causal graphs with feedback. To appear in *Uncertainty in Artificial Intelligence, Procedings, 12th Conference,* Morgan Kaufman, CA.

RICHARDSON, T. (1996a). A polynomial-time algorithm for deciding Markov equivalence of directed cyclic graphical models. In *Uncertainty in Artificial Intelligence, Proceedings, 12th Conference*, Morgan Kaufman, CA.

RICHARDSON, T. (1996b). A discovery algorithm for directed cyclic graphs. In *Uncertainty in Artificial Intelligence, Proceedings, 12th Conference,* Morgan Kaufman, CA.

SPIRTES, P., (1994), Conditional independence in directed cyclic graphical models representing feedback or mixtures. Philosophy, Methodology and Logic Technical Report 59, CMU

SPIRTES, P., C. GLYMOUR, R. SCHEINES (1993), Causation, Prediction and Search. Lecture Notes in Statistics, Springer-Verlag.