Conditional Independence in Directed Cyclic Graphical Models Representing Feedback or Mixtures

by

Peter Spirtes

September 1994

Report CMU-PHIL-59

Carnegie Mellon

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Conditional Independence in Directed Cyclic Graphical Models Representing Feedback or Mixtures

Peter Spirtes[1]

Department of Philosophy

Carnegie Mellon University

Pittsburgh, Pa. 15213

e-mail: ps7z@andrew.cmu.edu

[1] I thank C. Glymour, R. Scheines, M. Drudzel, H. Simon, T. Richardson, and C. Meek for helpful conversations. Research for this paper was supported by the National Science Foundation through grant 9102169 and the Navy Personnel Research and Development Center and the Office of Naval Research through contract number N00014-93-1-0568.

1. Introduction

The introduction of statistical models represented by directed acyclic graphs (DAGs) has proved fruitful in the construction of expert systems, in allowing efficient updating algorithms that take advantage of conditional independence relations (Pearl 1988, Spiegelhalter et. al. 1993), and in inferring causal structure from conditional independence relations (Spirtes and Glymour 1991, Spirtes, Glymour and Scheines 1993, Pearl and Verma 1991, Cooper 1992). As a framework for representing the combination of causal and statistical hypotheses, DAG models have shed light on a number of issues in statistics ranging from Simpson's Paradox to experimental design (Spirtes, Glymour and Scheines 1993). The relations of DAGs to statistical constraints, and the equivalence and distinguishability properties of DAG models, are now well understood, and their characterization and computation involves three properties connecting graphical structure and probability distributions: (i) a local directed Markov property, (ii) a global directed Markov property, (iii) and factorizations of joint densities according to the structure of a graph (Lauritizen, et al. 1990).

Recursive structural equation models are one kind of DAG model. However, non-recursive structural equation models are not DAG models, and are instead naturally represented by directed *cyclic* graphs in which a finite series of edges representing causal influence leads from a vertex representing a variable back to that same vertex. Such graphs have been used to model feedback systems in electrical engineering (Mason 1953, 1956), and to represent economic processes (Haavelmo 1943, Goldberger 1973). In contrast to the acyclic case, almost nothing general is known about how directed cyclic graphs (DCGs) represent conditional independence constraints, or about their equivalence or identifiability properties, or about characterizing classes of DCGs from conditional independence relations or other statistical constraints. This paper addresses the first of these problems, which is a prerequisite for the others. The issues turn on how the relations among properties (i), (ii) and (iii) essential to the acyclic case generalize~or more typically fail to generalize-to directed cyclic graphs and associated families of probability measures. It will be shown that when DCGs are interpreted by analogy with DAGs as representing functional dependencies with independently distributed noises or "error terms,[1]" the equivalence of the fundamental global and local Markov conditions characterise of DAGs no longer holds, even in linear systems, and in non-linear systems both Markov properties may fail. For linear systems associated with DCGs with independent errors or noises, a characterisation of conditional independence constraints is obtained, and it is shown that the result generalizes in a natural way to systems in which the error variables or noises are statistically dependent. For non-linear systems with independent errors a sufficient condition for conditional independence of variables in associated probability measures is obtained.

A second natural use of cyclic graphs is to represent mixtures in which in some subpopulations A causes B, and in other subpopulations B causes A. In section 5 it is shown how to construct cyclic graphs which represent the conditional independence relations in such mixtures.

The remainder of this paper is organized as follows: Section 2 defines relevant mathematical ideas and gives some necessary technical results on DAGs and DCGs. Section 3 obtains results for non-recursive linear structural equations models. Section 4 treats non-linear models of the same kind. Section 5 treats the use of cyclic graphs to represent mixtures. Except where they are necessary to the discussion, proofs of new results are given in the Appendix. Because the aim of this paper is to characterize conditional independence properties of formal structures implicit in various applied models, the discussions of motivation necessarily mix mathematical issues framed in graphical and probabilistic terms with a different terminology used in applied statistics. I have attempted to keep as closely as possible to the terminology in influential sources.

2. Directed Graphs

Sets of variables and defined terms are in boldface, and individual variables are in italics. A **directed graph** is an ordered pair of a finite set of vertices **V**, and a set of directed edges **E**. A directed edge from A to B is an ordered pair of distinct vertices $<A,B>$ in **V** in which A is the **tail** of the edge and B is the **head**; the edge is **out of** A and **into** B, and A is **parent** of B and B is a **child** of A. A sequence of edges $<E_1,...,E_n>$ in G is an **undirected path** if and only if there exists a sequence of vertices $<V_1,...,V_n>$ such that for $1 \leq i < n$ either $<V_i,V_{i+1}> = E_i$ or $<V_{i+1},V_i> = E_i$. A path U is **acyclic** if no vertex occurring on an edge in the path occurs more than once. A sequence of edges $<E_1,...,E_n>$ in G is a **directed path** if and only if there exists a sequence of vertices $<V_1,...,V_n>$ such that for $1 \leq i < n$ $<V_i,V_{i+1}> = E_i$. If there is an acyclic directed path from A to B or $B = A$ then A is an **ancestor** of B, and B is a **descendant** of A. A directed graph is **acyclic** if and only if it contains no directed cyclic paths.[2]

[2] An undirected path is often defined as a sequence of vertices rather than a sequence of edges. The two definitions are essentially equivalent for acyclic directed graphs, because a pair of vertices can be identified

A **directed acyclic graph** (DAG) G with a set of vertices **V** can be given two distinct interpretations. On the one hand, such graphs can be used to represent causal relations between variables, where an edge from A to B in G means that A is a direct cause of B relative to **V**. A **causal graph** is a DAG given such an interpretation. On the other hand, a DAG with a set of vertices **V** can also represent a set of probability measures over **V**. Following the terminology of Lauritzen et. al. (1990) say that a probability measure over a set of variables **V** satisfies the **local directed Markov property** for a DAG G with vertices **V** if and only if for every W in **V**, W is independent of **V\(Descendants(**W,G**)** $\cup$ **Parents(**W,G**))** given **Parents(**W,G**)**, where **Parents(**W,G**)** is the set of parents of W in G, and **Descendants(**W,G**)** is the set of descendants of W in G. A DAG G **represents** the set of probability measures which satisfy the local directed Markov property for G. The use of DAGs to simultaneously represent a set of causal hypotheses and a family of probability measures extends back to the path diagrams introduced by Sewell Wright (1934). Variants of DAG models were introduced in the 1980's in Wermuth(1980), Wermuth and Lauritzen(1983), Kiiveri, Speed, and Carlin (1984), Kiiveri and Speed(1982), and Pearl(1988). [3]

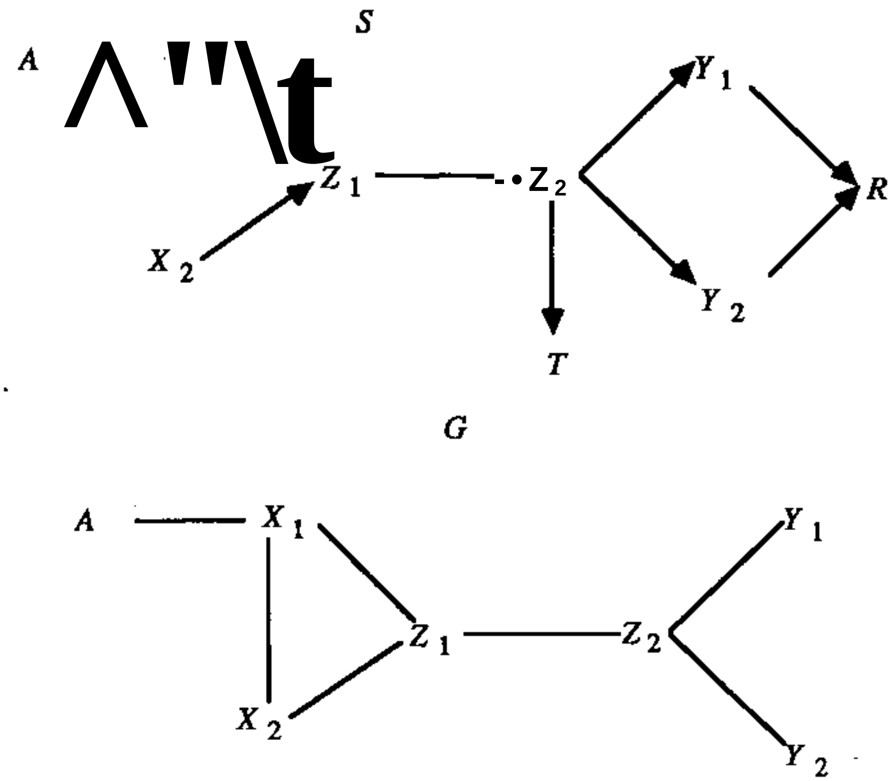
Lauritzen et. al. also define a **global directed Markov property** that is equivalent to the local directed Markov property for DAGs. Several preliminary notions are required. An **undirected graph** U is an ordered pair of a finite set of vertices **V**, and a set of undirected edges **E**. A sequence of undirected edges $<E_1,...,E_n>$ in U is an **undirected path** if and only if there exists a sequence of vertices $<V_1,...,V_n>$ such that for $1 \leq i < n$, V_i and V_{i+1} are adjacent in U. For a directed acyclic graph G, let V be a member of **An(X,**G**)** if and only if V is an ancestor of some member of **X** in G. Let G(**X**) be the subgraph of G that contains only vertices in **X**, with an edge from A to B in **X** if and only if there is an edge from A to B in G. If G is a directed graph, then $G^M(G)$ is an undirected graph with the same vertices as G, and a pair of vertices X and Y are adjacent in $G^M(G)$ if and only if either X and Y are adjacent in G, or they have a common child in G; $G^M(G)$ **moralizes** G. In an undirected graph G, **X** is **separated** from **Y** given **Z** if and only if

with a unique edge in the graph. However, a cyclic graph may contain more than one edge between a pair of vertices. In that case it is no longer possible to identify a pair of vertices with a unique edge.

[3]It is often the case that some further restrictions are placed on the set of distributions represented by a DAG. For example, one could also require the Minimality Condition, i.e. that for any distribution P represented by G, P does not satisfy the local directed Markov Condition for any proper subgraph of G. This condition and others are discussed in Pearl(1988) and Spirtes, Glymour, and Scheines(1993). We will not consider such further restrictions here.

every undirected path between a member of X and a member of Y contains a member of Z. G^M(G(An(X u Y u Z,G))) is the undirected graph that moralizes the subgraph of G that contains the vertices that are ancestors of X u Y u Z in G. If X, Y and Z are disjoint sets of variables, **X** and **Y** are **d-separated** given Z in a directed graph G just when X and Y are separated given Z in *G^M(G(An(X* u Y u Z,G))). Figure 1 illustrates how to form G^(G(An(X u Y u Z,G))), where X = *{$X_{iy}X_2$}>* *Y* = *{Y_hY_2}* and Z = {Zi,Z_2}.

(The relation I have called "d-separation" was introduced in Lauritzen et al. (1990) but was not called "d-separation" there. Since Lauritzen et al. (1990) proved that for directed acyclic graphs the graphical relation they defined is equivalent to the relation Pearl called "d-separation" in Pearl(1986), and the proof is readily extended to the cyclic case, "d-separation" will also be used to refer to the graphical relation introduced in Lauritzen et. al. 1990.)



G^M(G(An(X u Y u **Z,*G***)))

Figure 1

Now the definition: A probability measure over V satisfies the **global directed** Markov property for DAG G if and only if for any three disjoint sets of variables X, Y, and Z included in V, if X is d-separated from Y given Z, then X is independent of Y given Z. Lauritzen et. al. (1990) shows that the global and local directed Markov properties are equivalent in DAGs, even when the probability measures represented have no density functions. In section 2, it is shown that the local and global directed Markov properties are not equivalent for cyclic directed graphs.

The following lemmas relate the global directed Markov property to factorizations of a density function. Denote a density function over V by/(V), where for any subset X of V, *f(X)* denotes the marginal of/(V). If/(V) is the density function for a probability measure over a set of variables **V,** say that/(V) **factors according to directed graph** G with vertices V if and only if for every subset X of V, there exists non-negative functions *gv* such that

$$/(An(X,G)) = \prod_{VeAn(X,G)} g_V(V, \mathbf{Parents}(V,G))$$

The following result was proved in Lauritzen et. al. (1990).

Lemma 1: If V is a set of random variables with a probability measure *P* that has a density function/(V), then/(V) factors according to DAG G if and only if *P* satisfies the global directed Markov property for G.

As in the case of acyclic graphs, the existence of a factorization according to a cyclic directed graph G does entail that a measure satisfies the global directed Markov property for G. The proof given in Lauritzen et. al. (1990) for the acyclic case carries over essentially unchanged for the cyclic case.

Lemma 2: If V is a set of random variables with a probability measure *P* that has a density function/(V) and/[V) factors according to directed (cyclic or acyclic) graph G, then *P* satisfies the global directed Markov property for G.

However, unlike the case of acyclic graphs, if a probability measure over a set of variable V satisfies the global directed Markov property for cyclic graph G and has a density function /(V), it does not follow that /(V) factors according to G, as the following example, adapted from exercise 3.3 in Pearl(1988) shows. The probability measure in the following table (in which the final column contains the probability of the corresponding row of values for the random variables) satisfies the global directed Markov condition for the directed graph in figure 2, but does not factor according to that graph.

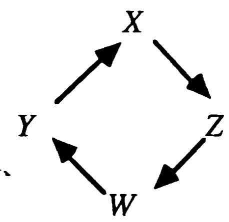


Figure 2

X	*Y*	*Z*	*W*	*P*
1	1	1	1	1/3
1	2	2	2	1/3
2	2	1	3	1/3
all other tuples				0

The following weaker result relating factorization of densities and the global directed Markov property does hold for both cyclic and acyclic directed graphs.

Lemma 3: If **V** is a set of random variables with a probability measure P that has a positive density function $f(\mathbf{V})$, and P satisfies the global directed Markov property for directed (cyclic or acyclic) graph G, then $f(\mathbf{V})$ factors according to G.

3. Non-recursive Linear Structural Equation Models

In a terminology that unavoidably mixes econometric and graphical ideas, the problem considered in this section is to investigate the generalization of the Markov properties to linear, non-recursive structural equation models, and, ultimately, to describe a fast algorithm that will decide correlation and partial correlation constraints entailed by such models. A secondary question concerns the equivalence of any linear structural equation model with correlated errors to a model with extra, latent, variables and uncorrelated errors. First I must relate the social scientific terminology to graphical representations, and clarify the questions.

Linear structural equation models with jointly independent error terms (which, adapting the terminology of Bollen (1989), will be referred to as linear SEMs) can be represented as directed graph models. In a linear SEM the random variables are divided into two disjoint sets, the error terms and the non-error terms. Corresponding to each non-error random variable *V* is a unique error term *By.* A linear SEM contains a set of linear equations in which each non-error random variable *V* is written as a linear function of other non-error random variables and *By.* A linear SEM also specifies a joint probability measure over the error terms. So, for example, the following is a linear SEM, where *a* and *b* are real constants, *Bx* >*By,* and *Bz* are jointly independent "error terms[11], and X, F, Z, are random variables:

$$X = a \times Y + \varepsilon_X$$
$$Y = b \times Z + \varepsilon_Y$$
$$Z = \varepsilon_Z$$

Bx> By_y and *Bz* are jointly independent with standard normal distributions

The directed graph of a linear SEM with uncorrelated errors is written with the convention that an edge does not appear if and only if the corresponding entry in the coefficient matrix is zero; the graph does not contain the error terms. Figure 3 is the DAG that represents the SEM shown above. A linear SEM is **recursive** if and only if its directed graph is acyclic.

Z ——-• 7———• X

Figure 3

Initially only linear SEMs in which the error terms are jointly independent will be considered, but in the linear case in an important sense nothing is lost by this restriction: a linear SEM with dependent errors generates the same restrictions on the covariance matrix as does some linear SEM with extra variables and independent errors. Further, such an SEM with extra variables can always be found with the same graphical structure on the original variables as obtain in the original graph.

A directed graph containing disjoint sets of variables **X, Y,** and **Z linearly entails** that X is independent of Y given Z if and only if X is independent of Y given Z for all values of the linear coefficients not fixed at zero, and all probability measures over the error variables in which they are jointly independent and have positive variances. Let *pxjx* be the partial correlation of *X* and *Y* given Z. A directed graph containing *X*, Y_y and Z, where *X* * *Y* and *X* and *Y* are not in **Z, linearly entails** that *pxjjz* = 0 if and only *pxj.z* = 0 f°r all values of the linear coefficients not fixed at zero, and all probability measures over the error variables in which they are uncorrelated, have positive variances, and in which $\rho_{X,Y.Z}$ is defined. It follows from Kiiveri and Speed (1982) that any probability measure over non-error terms derived from a linear, recursive SEM with jointly independent error terms and directed graph G, satisfies the local directed Markov property for G. One can therefore apply d-separation to the DAG in a linear, recursive SEM to compute the conditional independencies and zero partial correlations it linearly entails. The d-separation relation provides a polynomial (in the number of vertices) time algorithm for deciding whether a given vanishing partial correlation is linearly entailed by a DAG.

Linear non-recursive structural equation models (linear SEMs) are commonly used in the econometrics literature to represent feedback processes that have reached equilibrium.[4] Corresponding to a set of non-recursive linear equations is a cyclic graph, as the following example from Whittaker(1990) illustrates.[5]

$$X_1 = \varepsilon_{X_1}$$
$$X_2 = \varepsilon_{X_2}$$
$$X_3 = \beta_{31} \times X_1 + \beta_{34} \times X_4 + \varepsilon_{X_3}$$
$$X_4 = \beta_{42} \times X_2 + \beta_{43} \times X_3 + \varepsilon_{X_4}$$
$\varepsilon_{X_1}, \varepsilon_{X_2}, \varepsilon_{X_3}, \varepsilon_{X_4},$ are jointly independent and normally distributed

$$\begin{tikzcd} X_1 \arrow[r] & X_3 \arrow[d, shift right] \\ X_2 \arrow[r] & X_4 \arrow[u, shift right] \end{tikzcd}$$

Figure 4

[4]Cox and Wermuth(1993), Wermuth and Lauritzen(1990) and (indirectly) Frydenberg(1990) consider a class of non-recursive linear models they call *block recursive*. The block recursive models overlap the class of SEMs, but they are neither properly included in that class, nor properly include it. Frydenberg (1990) presents necessary and sufficient conditions for the equivalence of two block recursive models.

[5]Note that this use of cyclic directed graphs to represent feedback processes represents an extension of the causal interpretation of directed graphs. The causal structure corresponding to figure 4 is described by an infinite acyclic directed graph containing each variable indexed by time. The cyclic graph can be viewed as a compact representation of such a causal graph. I am indebted to C. Glymour for pointing out that the local Markov condition fails in Whittaker's model. Indeed, there is *no* acyclic graph (even with additional variables) that linearly entails all and only conditional independence relations linearly entailed by figure 4, although Thomas Richardson has pointed out that the directed cyclic graph of figure 4 is equivalent to one in which in the edges from X_1 to X_3 and X_2 to X_4 are replaced, respectively, by edges from X_1 to X_4 and from X_2 to X_3.

Theorem 1: In a directed (cyclic or acyclic) graph G containing disjoint sets of variables X, Y and Z, if X is d-separated from Y given Z then G linearly entails that X is independent of Y given Z.

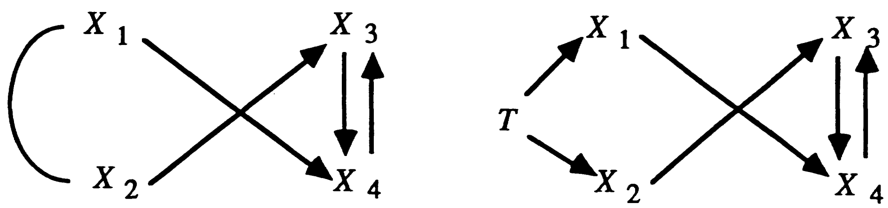
Theorem 2: In a directed (cyclic or acyclic) graph *G* containing disjoint sets of variables X, Y and Z, if X is not d-separated from Y given Z then *G* does not linearly entail that X is independent of Y given Z.

Applying Theorems 1 and 2 to the directed graph in figure 4, two conditional independence relations are entailed: *X* is independent of *X%* and *X* is independent of *Xi* given X3 and *X4.* In DAGs the global directed Markov property entails the local directed Markov property, because a variable *V* is d-separated from its non-parental non-descendants given its parents. This is not always the case in cyclic graphs. For example, in figure 4, *X4* is not d-separated from its non-parental non-descendant *X* given its parents *X2* and X3, so the local directed Markov property does not hold.

Theorem 3: In a (cyclic or acyclic) directed graph *G* containing X_y *Y* and Z, where *X* * *Y* and Z does not contain *X* or 7, *X* is d-separated from *Y* given Z if and only if G linearly entails that px,r.z = 0.

As in the acyclic case, d-separation provides a polynomial time procedure for deciding whether cyclic graphs entail a conditional independence or vanishing partial correlation.

Theorem 3 can also be used to calculate the zero partial correlations entailed for all values of the linear coefficient not fixed at zero, and all probability measures over the error variables where the variances are positive and the partial correlations exist, even if the error terms are correlated. If *ex* and *ey* are not independent in linear SEM L, there is a linear SEM *V* with independent error terms such that the marginal probability measure of *L* over the variables in *L* has the same covariance matrix as L. Form the graph $G^{\%}$ of *L* from the graph *G* of *L* in the following way. Add a latent variable *T* to G, and add edges from *T \oX* and *Y.* In *L* modify the equation for *X* by making it a linear functions of the parents of *X* (including *T)* in *G* and replace *ex* by $e^{\%}x\backslash$ modify the equation for *Y* in an analogous way. There always exist linear coefficients and probability measures over *T* and the new error terms such that the marginal covariance matrix for *V* is equal to the covariance matrix of L, and e'x and €*y* are independent. The process can be repeated for each pair of variables with correlated errors in L. Hence the zero partial correlations entailed by L can be derived by applying Theorem 3 to the graph of L'. Figure 5 illustrates this process. The set of variables **V** in the graph on the left is $\{X_1,X_2,X_3,X_4\}$. The graph on the left correlates the errors between X_1 and X_2 (indicated by the undirected edge between them.) The graph on the right has no correlated errors, but does have a latent variable T that is a parent of X_1 and X_2. The two graphs linearly entail the same zero partial correlations involving only variables in **V** (in this case they both entail no non-trivial zero partial correlations).



Graph with Correlated Error | **Graph without Correlated Error and Same Partial Correlations Over V**

Graph with Correlated Error	Graph without Correlated Error
$X_3 = a \times X_2 + b \times X_4 + \varepsilon_3$	$X_3 = a \times X_2 + b \times X_4 + \varepsilon_3$
$X_4 = c \times X_1 + d \times X_3 + \varepsilon_4$	$X_4 = c \times X_1 + d \times X_3 + \varepsilon_4$
$X_1 = \varepsilon_1$	$X_1 = e \times T + \varepsilon'_1$
$X_2 = \varepsilon_2$	$X_2 = f \times T + \varepsilon'_2$
ε_1 and ε_2 correlated	ε'_1 and ε'_2 uncorrelated

Figure 5

3. Non-linear Structural Equation Models

In a SEM, the equations relating a given variable to other variables and a unique error term need not be linear. In a SEM the random variables are divided into two disjoint sets, the error terms and the non-error terms. Corresponding to each non-error random variable V is a unique error term ε_V. It will be assumed that the error terms are jointly independent. A SEM contains a set of equations in which each non-error random variable V is written as a measureable function of other non-error random variables and ε_V. The convention is that in the directed graph of a SEM there is an edge from A to B if and only if A is an argument in the function for B. It will be assumed that density functions exist for both the probability measure over the error terms and the probability measure over the non-error terms, that each non-error term V can also be written as a function of the error terms of its ancestors in G, that each ε_V is a function of V and its parents in G (which will be the case if the errors are additive or multiplicative), and that the Jacobean of the transformation between the error terms and the non-error terms is well-defined. Call such a set of equations and its associated graph a **pseudo-indeterministic** SEM (because the equations are actually deterministic if the unmeasured error terms are included, but appear indeterministic when the error terms are not measured.) A directed graph G **pseudo-indeterministically entails** that **X** is independent of **Y** given **Z** if and only if in every pseudo-indeterministic SEM with graph G and jointly independent errors, **X** is independent of **Y** given **Z**.

This section establishes that d-separation again provides a fast algorithm for deciding whether a DAG pseudo-indeterministically entails a conditional independence relation, but in a DCG d-separation may not pseudo-indeterministically entails a conditional independence relation. Instead, a different condition, yielding a polynomial time algorithm, is found to suffice for a cyclic directed graph to pseudo-indeterministically entail a conditional independence relation.

By Theorem 2, d-separation is a necessary condition for a conditional independence relation to be pseudo-indeterministically entailed by a directed graph. The following remarks show d-separation is also a sufficient condition for pseudo-indeterministic entailment in acyclic directed graphs, but not for cyclic directed graphs.

Theorem 4: If G is a DAG containing disjoint sets of variables **X**, **Y** and **Z**, **X** is d-separated from **Y** given **Z** if and only if G pseudo-indeterministically entails that **X** is independent of **Y** given **Z**.

It is instructive to see why the proof that a DAG G pseudo-indeterministically entails that **X** is d-separated from **Y** given **Z** if and only L entails that **X** is independent of **Y** given **Z** breaks down in the case of cyclic directed graphs. In both cyclic and acyclic directed graphs it follows that

$$f(\mathbf{An}(\mathbf{X},G)) = \prod_{X \in \mathbf{An}(\mathbf{X},G)} f(g_X(X,\mathbf{Parents}(X,G)) \times |J|$$

where J is the Jacobian of the transformation from the non-error terms to the error terms, and each g_X is a non-negative function. However, in a DAG, the absolute value of the Jacobian of the transformation is a single term consisting of the product of the terms along the diagonal of the transformation matrix:

$$|J| = \left|\prod_{V\in An(X,G)} \frac{de_V}{\sim dV}\right| = \prod_{V\in An(X,G)} \left|\frac{\partial \varepsilon_V}{dV}\right| = \prod_{V\in An(X,G)} m_V(V, \mathbf{Parents}(V,G))$$

(This is because for an acyclic graph the transformation matrix can be arranged so that it is lower triangular.) Each term *\dsy I dV\\s* some non-negative function *my* of *V* and its parents, because *ey* is a function of *V* and its parents. Hence by lemma 1, if X and Y are d-separated given Z, then X and Y are independent given Z.

However, if G is s cyclic directed graph, the Jacobian of the transformation is not in general a single term, but is the sum of several terms. The absolute value of the Jacobian can be expressed as

$$|J| = \left|\sum_i \prod_{V\in An(X,G)} m_{i,V}(V, \mathbf{Parents}(V,G))\right|$$

There is no reason to expect, however, that in the probability measure with density formed from the sum, X is independent of Y given Z. The global directed Markov property thus fails.

The following example gives a concrete illustration that there is a cyclic graph G in which *X* is d-separated from *Y* given {Z,W}> but G does not pseudo-indeterminstically entail that *X* is independent of *Y* given {Z,W}.

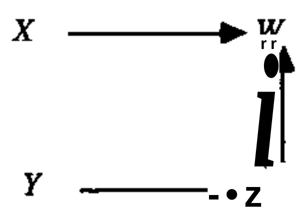


Figure 6: Graph *G*

$$X = e_x$$
$$Y = \varepsilon_Y$$
$$Z = WxY + \varepsilon_Z$$
$$W = Z \times X + \varepsilon_W$$

ex, £y £Z» e*W* with independent standard normal distributions

The transformation from *ex. By, Bz, By/* to X, F, Z, iy is 1-1 except where *ex* x *By* =1 because

$$X = \varepsilon_X$$
$$Y = \varepsilon_Y$$
$$Z = \frac{\varepsilon_W \times \varepsilon_Y + \varepsilon_Z}{1-(\varepsilon_X \times \varepsilon_Y)}$$
$$W = \frac{\varepsilon_Z \times \varepsilon_X + \varepsilon_W}{1-(\varepsilon_X \times \varepsilon_Y)}$$

The Jacobean of the transformation from the e's is 1/(1 - *x* x y). Hence, transforming the joint normal density of the *e's* yields

$$f(x,y,z,w) =$$
$$\frac{1}{4\pi^2}\left(\exp\left(-\frac{1}{2}(x^2 + y^2 + (z-wxy)^2 + (w-zxx)^2)\right)\right) \times \left|\frac{1}{1-(x \times y)}\right|$$

X is not independent of *Y* given *{Z_yW}* in a probability measure with this density because it is not possible to factor the density into a product of terms where none of the terms contains both *X* and *Y.*

However, it is possible to modify the graphical representation of the functional relations in such a way that d-separation applied to the new graph does entail conditional independence. In a directed graph G, a cycle is a cyclic directed path, in which each vertex occurs on exactly two edges in the path. A set of cycles C is a cyclegroup if and only if it is a smallest set of cycles such that for each cycle *C* in C, C contains the transitive closure of all of the cycles intersecting Ci, i.e. it contains all of the cycles that intersect Ci, all of the cycles that intersect cycles that intersect Ci, etc. For example, in figure 7, there are two distinct cyclegroups: the first is {Ci,C2,C3}, and the second is $\{C_4, C_5\}$.

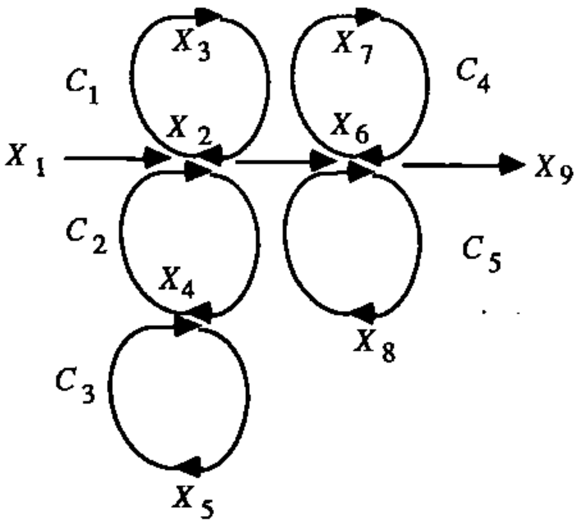


Figure 7

Form the **collapsed graph** G^f from G by the following operations on each cyclegroup:

1. remove all of the edges between vertices in cycles in the cyclegroup;
2. arbitrarily number the vertices in the cyclegroup;
3. add an edge from each lower number vertex to each higher number vertex;
4. for each parent A of a member of the cyclegroup that is not itself in the cyclegroup, add an edge from *A* to each member of the cyclegroup.

(The procedure does not define a unique collapsed graph due to the arbitrariness of the numbering, but since all of the collapsed graphs share the same d-separation relations, it does not matter.) Note that even if G is a cyclic graph, the collapsed graph is acyclic. The collapsed graph can be generated in polynomial time.

Theorem 5: If *G* is a directed graph (cyclic or acyclic), collapsed graph G^1 contains disjoint sets of variables X, Y and Z, and X is d-separated from Y given Z in G^1 then G pseudo-indeterministically entails that X is independent of Y given Z.

A collapsed graph for the graph in figure 7 is shown in figure 8a, and a collapsed graph for the graph in figure 4 is shown in figure 8b.

I do not know whether the follow conjecture holds:

Conjecture: Let G (cyclic or acyclic) have collapsed graph G' containing disjoint sets of variables **X**, **Y** and **Z**. If G pseudo-indeterministically entails that **X** is independent of **Y** given **Z**, then in G', **X** is d-separated from **Y** given **Z**.

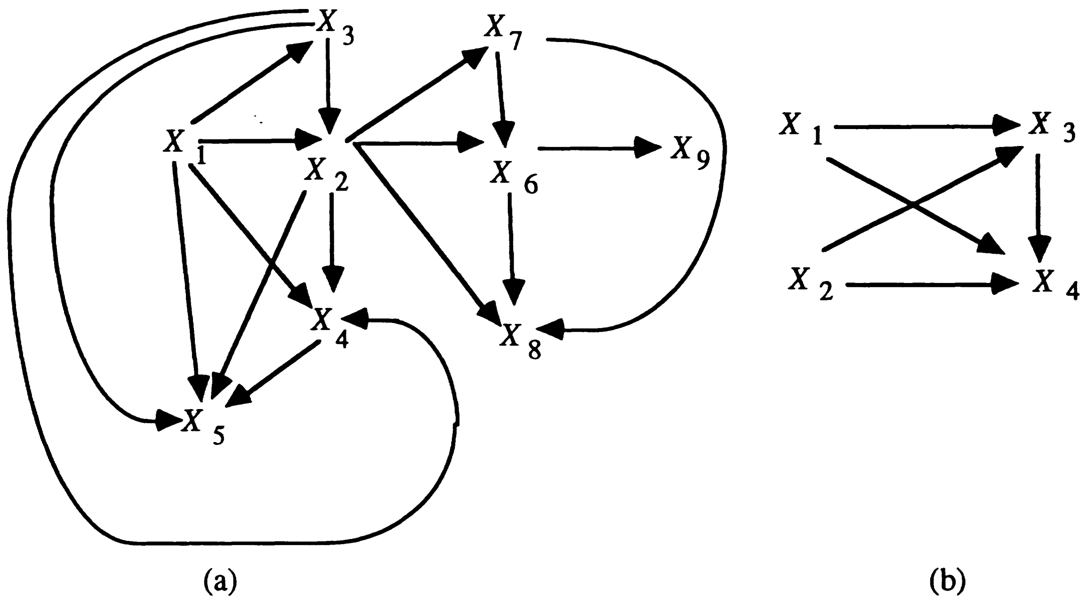


Figure 8

4. Mixtures

Sometimes the most reasonable hypothesis about real populations is that distinct causal processes are at work in distinct subpopulations. Suppose that for each subpopulation i, the causal processes at work in that population are represented by a directed graph G_i satisfying the global directed Markov property for the probability measure corresponding (perhaps ideally) to the subpopulation. Is it still possible to construct a graph G_{Mix} that represents both the combination of causal relations in the entire population and for which the mixed probability measure satisfies the global Markov property? In this section an affirmative answer is found.

Suppose that each G_i contains the same set of variables **V** and represents a probability measure that factors according to G_i, i.e.

$$f_i(\mathbf{An}(\mathbf{X},G_i)) = \prod_{X \in \mathbf{An}(\mathbf{X},G_i)} g_{i,X}(X, \mathbf{Parents}(X,G_i))$$

where each $g_{i,X}$ is a non-negative function. By lemmas 1 through 3, any probability measure with a density function represented by an acyclic graph has this property, and any probability measure with a positive density function represented by a cyclic graph has this property. For a given factorization of this form and directed graph G_i, each vertex V in G_i is associated with the parameter $g_{i,V}$ that represents a term in the factorization of the density function for the i^{th} population. Form a directed graph G_{Mix} that represents the mixture probability measure in the following way:

1. Let $\mathbf{V}_{\mathbf{Mix}} = \mathbf{V} \cup \{T\}$, where T is a variable not in $\mathbf{V}$, which takes on value i in the i^{th} subpopulation.
2. For each pair of variables A and B in $\mathbf{V}$, there is an edge from A to B in G_{Mix} if and only if there is an edge from A to B in G_i for some i.
3. If there exists a V in $\mathbf{V}$, and i and j such that $g_{i,V} \neq g_{j,V}$ then add an edge from T to V.

Theorem 6: If $P_{Mix}(\mathbf{V})$ is a mixture of probability measures, each of which factors according to directed graph G_i over $\mathbf{V}$, $P_{Mix}(\mathbf{V})$ satisfies the global directed Markov property for G_{Mix}.

Figure 9 shows G_{Mix} for a population consisting of two subpopulations with graphs G_1 and G_2 respectively, where the joint distribution of X_1 and X_2 are the same in each subpopulation, and the conditional distribution of X_5 on X_4 is the same in each subpopulation.

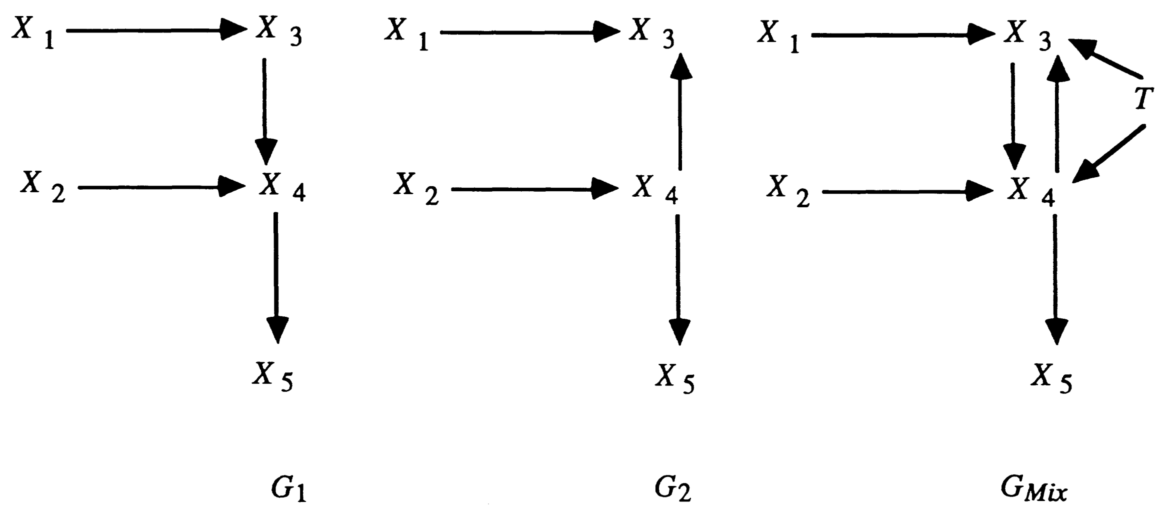


Figure 9

Note that the independence relations entailed by G_{Mix} are not the same as the intersection of the conditional independence relations in the two subpopulations, nor is there any directed acyclic graph which entails the same conditional independence relations as G_{Mix}. In this case, *X* and *X2* are independent in the mixture because they are independent in both subpopulations, and the joint distribution of *X* and *X2* is the same in both subpopulations. *X* and *X2* are independent conditional on X3 and *X4* in both subpopulations, but not in the mixture. (However, *X* and *X2* are independent conditional on X3, *X4* and *T* in the mixture.) In contast, *X* is independent of *X5* given *X4* in both subpopulations and in the mixture, eyenthough the joint distribution of *X\> X$* and *X4* is not the same in both subpopulations. There are also conditional independence relations in some subpopulations, but not in the mixture, which are not entailed by the graph. For example *X* is independent of *X2* given X3 in the first subpopulation, but not in the mixture, or in the second subpopulation.

5. Conclusion

These results raise a number of interesting questions whose answers may be of practical importance. Under what conditions, for example, are there results about conditional independence comparable to the equivalence of vanishing partial correlations in models with dependent errors and latent variable models with independent errors? There are polynomial algorithms (Verma and Pearl 1990, Frydenberg 1990) for determining when two arbitrary directed acyclic graphs entail the same set of conditional independence relations. Is there a polynomial algorithm for determining when two arbitrary directed graphs (cyclic or acyclic) linearly entail the same set of conditional independence relations? There are polynomial algorithms (Spirtes and Verma 1992) for determining when two arbitrary directed acyclic graphs entail the same set of conditional independence relations over a common subset of variable O. Is there a polynomial algorithm for determining when two arbitrary directed graphs (cyclic or acyclic) linearly entail the same set of conditional independence relations over a common subset of variables O? Assuming Markov properties hold and completely characterize the conditional independence facts in the probability measures considered, there are correct polynomial algorithms for inferring features of (sparse) directed acyclic graphs from a probability measure when there are no latent common causes (see Spirtes and Glymour 1991, Cooper and Herskovitz 1992). Are there comparable correct, polynomial algorithms for inferring features of directed graphs (cyclic or acyclic) from a probability measure when there are no latent common causes? There are similarly correct, but not polynomial, algorithms for inferring features of directed acyclic graphs from a probability measure even when there may be latent common causes (see Spirtes, 1992 and Spirtes, Glymour and Scheines 1993). Are there comparable algorithms for inferring features of directed graphs (cyclic or acyclic) from a probability measure even when there may be latent common causes?

Appendix

Lemma 3: If **V** is a set of random variables with a probability measure P that has a positive density function $f(\mathbf{V})$, and P satisfies the global directed Markov property for directed (cyclic or acyclic) graph G, then $f(\mathbf{V})$ factors according to G.

Proof. Assume that probability measure over **V** satisfies the global directed Markov property for directed (cyclic or acyclic) graph G. I will now show that for any disjoint sets of variables **R**, **S**, and **T** included in $\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)$, if **R** and **S** are separated given **T** in $G^M(G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)))$, then **R** and **S** are independent given **T**. If **R**, **S**, and **T** are included in $\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)$, then $\mathbf{An}(\mathbf{R} \cup \mathbf{S} \cup \mathbf{T},G)$ is included in $\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)$. Any pair of vertices A and B adjacent in $G^M(G(\mathbf{An}(\mathbf{R} \cup \mathbf{S} \cup \mathbf{T},G)))$ is also adjacent in $G^M(G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)))$ because $G(\mathbf{An}(\mathbf{R} \cup \mathbf{S} \cup \mathbf{T},G))$ is a subgraph of $G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G))$. Hence $G^M(G(\mathbf{An}(\mathbf{R} \cup \mathbf{S} \cup \mathbf{T},G)))$ is a subgraph of $G^M(G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)))$. It follows that if **R** and **S** are separated given **T** in $G^M(G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)))$ they are also separated in $G^M(G(\mathbf{An}(\mathbf{R} \cup \mathbf{S} \cup \mathbf{T},G)))$. But by the global directed Markov property, if **R** and **S** are separated given **T** in $G^M(G(\mathbf{An}(\mathbf{R} \cup \mathbf{S} \cup \mathbf{T},G)))$ then **R** and **S** are independent given **T**. It follows from the Hammersly-Clifford Theorem (see Lauritzen et. al. 1990) that the density function $f(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G))$ can be written as a product of non-negative functions of cliques[6] in $G^M(G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)))$. Since each set of parents of a vertex in $\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)$ is a clique in $G^M(G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)))$,

$$f(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)) = \prod_{V \in \mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G)} g_V(V,\mathbf{Parents}(V,G))$$

where each g_V is a positive function, i.e., the density function factors according to G. $\therefore$

Theorem 1: In a directed (cyclic or acyclic) graph G containing disjoint sets of variables **X**, **Y** and **Z**, if **X** is d-separated from **Y** given **Z** then G linearly entails that **X** is independent of **Y** given **Z**.

[6]Here "clique" refers to any compeletely connected subgraph. Some authors use "clique" to refer only to maximal completely connected sugraphs.

Proof. Let $\mathbf{Err(X)}$ be the set of error terms corresponding to a set of non-error variables $\mathbf{X}$. In order to distinguish the density function for the set $\mathbf{V}$ of non-error variables from the density function for the error variables I will use $f_{\mathbf{V}}$ to represent the density function (including marginal densities) for the former and $f_{\mathbf{Err}}$ to represent the density function of the latter. If $\mathbf{V}$ is the set of variables in G, then by hypothesis,

$$f_{\mathbf{Err}}(\mathbf{Err(V)}) = \prod_{\varepsilon \in \mathbf{Err(V)}} f_{\mathbf{Err}}(\varepsilon)$$

It is possible to integrate out the error terms not in $\mathbf{Err(An(X,}G\mathbf{))}$ and obtain

$$f_{\mathbf{Err}}(\mathbf{Err(An(X,}G\mathbf{)))} = \prod_{\varepsilon \in \mathbf{Err(An(X,}G\mathbf{))}} f_{\mathbf{Err}}(\varepsilon)$$

Because for each variable X in $\mathbf{V}$, X is a linear function of its parents in G plus a unique error term ε_X, it follows that ε_X is a linear function g_X of X and the parents of X in G. Hence $\mathbf{Err(An(X,}G\mathbf{))}$ is a function of $\mathbf{An(X,}G\mathbf{)}$. Following Haavelmo(1943) it is possible to derive the density function for the set of variables $\mathbf{An(X,}G\mathbf{)}$ by replacing each ε_X in $f_{\mathbf{Err}}(\varepsilon_X)$ by $g_X(X,\mathbf{Parents}(X))$ and multiplying by the absolute value of the Jacobean:

$$f_{\mathbf{V}}(\mathbf{An(X,}G\mathbf{))} = \prod_{X \in \mathbf{An(X,}G\mathbf{)}} f_{\mathbf{Err}}(g_X(X,\mathbf{Parents}(X,G))) \times |J|$$

where J is the Jacobian of the transformation. Because the transformation is linear, the Jacobian is a constant. All of the terms in the multiplication are non-negative because they are either a density function or a positive constant. It follows from lemma 1 that if $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$ then $\mathbf{X}$ and $\mathbf{Y}$ are independent given $\mathbf{Z}$. $\therefore$

Lemma 4: In a directed graph G with vertices $\mathbf{V}$, if $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ are disjoint subsets of $\mathbf{V}$, and $\mathbf{X}$ is d-connected to $\mathbf{Y}$ given $\mathbf{Z}$ in G, then $\mathbf{X}$ is d-connected to $\mathbf{Y}$ given $\mathbf{Z}$ in an directed acyclic subgraph of G.

Proof. I will use the sense of d-connection defined in Pearl(1988) which Lauritzen et. al. (1990) proved equivalent to their sense of d-connection for acyclic graphs. The proof of the equivalence given by Lauritzen et. al can easily be extended to cyclic graphs. Vertex X is a **collider** on an acyclic undirected path U in directed graph G if and only if there are two edges on U that are directed into X. According to Pearl's definition, for three disjoint sets $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ and $\mathbf{Y}$ are **d-separated** given $\mathbf{Z}$ in G if and only if there is no acyclic undirected path U from a member of X to a member of Y such that every non-collider on U is not in Z, and every collider on U has a descendant in Z. For three disjoint sets **X, Y,** and **Z, X** and **Y** are **d-connected** given Z in G if and only if X and Y are not **d-separated** given Z.

Suppose that U is an undirected path that d-connects X and Y given Z, and C is a collider on U. Let *length(C₉Z)* be 0 if C is a member of Z, or the length of a shortest directed path from C to a member of Z if C is not in Z. Let *size(U)* equal the number of colliders on U plus the sum over all colliders C on U of *length(C,Z).* U is a **minimal path** that d-connects X and Y given Z, if there is no other path U^l that d-connects X and 7 given Z such that *size(U%) < size(U).* If there is a path that d-connects X and Y given Z there is at least one minimal path that d-connects X and Y given Z.

Suppose X is d-connected to Y given Z, Then for some X in X and Y in Y, X is d-connected to 7 given Z by some minimal path U in G. First I will show that no shortest acyclic directed path £>/ from a collider C/ on t/ to a member of Z intersects £/ except at C/. Suppose this is false. I will show that it follows that there is a path $U^{\%}$ that d-connects X and Y given Z such that *size(JJ*) < size(U),* contrary to the assumption that U is minimal. See figure 10.

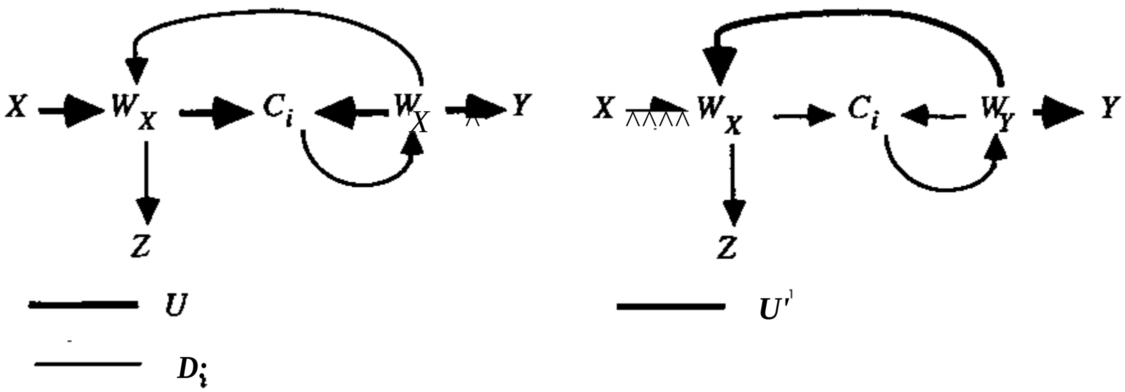


Figure 10

Form the path *If* in the following way. If D,·intersects U at a vertex other than C/ then let Wx be the vertex on D,- and U that is closest to X on £/, and *Wy* be the vertex on £>/ and £/ that is closest to Y on [/. Suppose without loss of generality that *Wx* is after *Wy* on £>/. Let £/' be the concatenation of *U(X,W$_X$), D&WyWx),* and *U(Wy,Y)* (where f/(X'',Wx) denotes the subpath of U between X and *Wx-)* It is now easy to show that *IT* d-connects X and Y given Z, and *size(U') < size(U)* because *U'* contains no more colliders than U and a shortest directed path from *Wx* to a member of Z is shorter than £>/. Hence *U* is not minimal, contrary to the assumption. If *D[* intersects *U* at only one vertex other than C; the proof is similar.

Next, I will show that if *U* is minimal, then it does not contain a pair of colliders *C* and *D* such that a shortest directed path from *C* to a member of Z intersects a shortest path from *D* to a member of Z. Suppose, this is false. See figure 11.

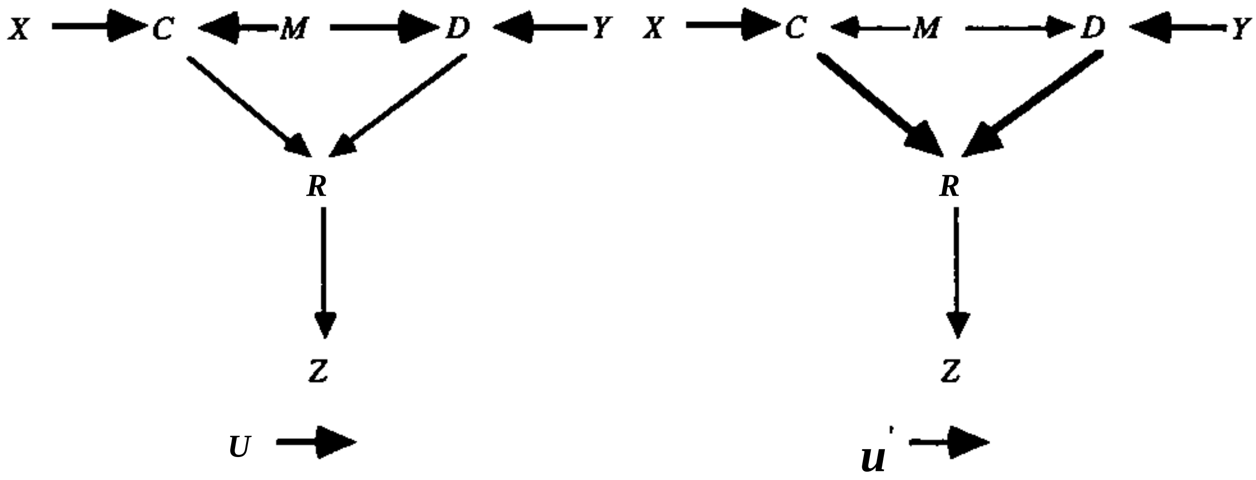


Figure 11

Let *D* be a shortest directed acyclic path from C to a member of Z that intersects Z>2> a shortest directed acyclic path from *D* to a member of Z. Let the vertex on *D* closest to *C* that is also on *Di* be *R*. Let t/[1] be the concatenation of *U(X,C)*, *D\{CJi)*, £>2(C>#), and *U{YJ))*. It is now easy to show thatf/[1] d-connects *X* and 7 given Z and *size{U%) < size(U)* because *U%* contains fewer colliders than *U*. Hence *U* is not minimal, contrary to the assumption. If *R* = C or *R* = *D*, the proof is similar.

For each collider *C* on a minimal path *U* that d-connects X and *Y* given Z, a shortest directed path from *C* to a member of Z does not intersect £/ except at C, and does not intersect a shortest directed path from any other collider *D* to a member of Z. It follows that the subgraph consisting of *U* and a shortest directed acyclic path from each collider on *U* to a member of Z is acyclic. .\

Theorem 2: In a directed (cyclic or acyclic) graph G containing disjoint sets of variables X, Y and Z, if X is not d-separated from Y given Z then G does not linearly entail that X is independent of Y given Z.

Proof. Suppose then that X is not d-separated from Y given Z. By lemma 4, if X is not d-separated from Y given Z in a cyclic graph G, then there is some acyclic subgraph G^f of G in which X is not d-separated from Y given Z. Geiger and Pearl (1988) have shown that if X is not d-separated from Y given Z in a DAG, then there is some probability measure represented by the DAG in which X is not independent of Y given Z, and it has been shown (Spirtes, Glymour and Scheines 1993) that there is in particular a linear normal distribution *P* in which X is not independent of Y given Z. If /^satisfies the global directed Markov property for G^1 it also satisfies it for G because every d-connecting path in G^1 is a d-connecting path in G. Hence there is some linear normal distribution represented by G in which X is not independent of Y given **Z.** /.

Theorem 3: In a (cyclic or acyclic) directed graph G containing *X*> *Y* and Z, where *X* * *Y* and Z does not contain *X* or Y_y *X* is d-separated from *Y* given Z if and only if G linearly entails that px,y.z = 0.

Proof. (This proof for cyclic or acyclic graphs is based on the proof for acyclic graphs in Verma and Pearl 1990.) Suppose that *X* and *Y* are d-separated given Z in G. Let *L* be a linear RSEM with the same linear coefficients and covariance marix among the error variables as L, but with normally distributed error variables. In *L* the error terms are jointly independent, and hence *X* is independent of *Y* given Z. It follows that *pxjx* = 0 in *L* But the correlation matrix depends only upon the covariance matrix among the error variables and the linear coefficients. Hence *L* and *L* have the same correlation matrix over the non-error variables, and the same partial correlations. It follows that *pxjz* = 0 inL.

Suppose that *X* and *Y* are d-connected given Z in G. Then by lemma 4, *X* and *Y* are d-connected given Z in some acyclic subgraph G^1 of G. In Spirtes, Glymour and Scheines (1993) it is shown that if *X* and *Y* are d-connected given Z in some DAG *G* then there is a linear RSEM with uncorrelated errors and a probability measure represented by G^1 in which *pxj.Z* * 0. Since G^1 is a subgraph of G, the probability measure is also represented by G. /.

Note that for an SEM with graph G, if *V* *X_y then de_v *I* <?Xis non-zero only if there is an edge from *X* to *V* in G (because *By* is a function only of *V* and *Vs* parents in G.) Associate

with each non-zero partial derivative $\partial \varepsilon_V / \partial X$ the edge from X to V in G. A product of partial derivatives form a **loop** in G if and only if the corresponding edges form a cycle in G. Two loops **intersect** if and only if their corresponding cycles intersect.

Let $J_{\mathbf{Err(V)->V}}$ be the Jacobean of the transformation from **Err(V)** to **V**, and $J_{\mathbf{V->Err(V)}}$ be the Jacobean of the transformation from **V** to **Err(V)**. We will say that a variable W occurs in a partial derivative in the Jacobean $J_{\mathbf{V->Err(V)}}$ if W occurs in the denominator, or ε_W occurs in the numerator. For example, we say that W and X occur in $\partial \varepsilon_W / \partial X$ A product of partial derivatives S occurring in a term T in $J_{\mathbf{Err(V)->V}}$ is **minimally sufficient** in T if for each variable occurring in S, all of its occurrences in T are in S, and no subset of S has this property. For example, in

$$\frac{\partial \varepsilon_W}{\partial X} \times \frac{\partial \varepsilon_X}{\partial Y} \times \frac{\partial \varepsilon_Y}{\partial W} \times \frac{\partial \varepsilon_U}{\partial U} \times \frac{\partial \varepsilon_V}{\partial V}$$

the three minimally sufficient products are

$$\frac{\partial \varepsilon_W}{\partial X} \times \frac{\partial \varepsilon_X}{\partial Y} \times \frac{\partial \varepsilon_Y}{\partial W}, \quad \frac{\partial \varepsilon_U}{\partial U}, \quad \text{and } \frac{\partial \varepsilon_V}{\partial V}$$

$\left|J_{\mathbf{Err(V)->V}}\right|$ is equal to $\left|1 / J_{\mathbf{V->Err(V)}}\right|$, but it turns out to simplify the proofs if at intermediate stages $\left|J_{\mathbf{V->Err(V)}}\right|$ is used instead of $\left|J_{\mathbf{Err(V)->V}}\right|$. $J_{\mathbf{V->Err(V)}}$ is the determinant of a matrix in which the element in the i^{th} row and j^{th} column is $\partial \varepsilon_{V_i} / \partial V_j$.

X is an **ancestral set** for a directed graph G with vertices **V** if and only if **X** = **An**(**Y**,G) for some **Y** included in **V**.

Theorem 4: If G is a DAG containing disjoint sets of variables **X**, **Y** and **Z**, **X** is d-separated from **Y** given **Z** if and only if G pseudo-indeterministically entails that **X** is independent of **Y** given **Z**.

Proof. The first part of the proof is essentially the same as that of Theorems 1 and 2, and shows that

$$f(\mathbf{An}(\mathbf{X},G)) = \prod_{X \in \mathbf{An}(\mathbf{X},G)} f(g_X(X, \mathbf{Parents}(X,G)) \times |J|$$

In an acyclic graph, the Jacobian of the transformation is a single term consisting of the product of the terms along the diagonal of the transformation matrix:

$$|J| = \left| \prod_{V \in \mathbf{An}(\mathbf{X},G)} \frac{\partial \varepsilon_V}{\partial V} \right| = \prod_{V \in \mathbf{An}(\mathbf{X},G)} \left| \frac{\partial \varepsilon_V}{\partial V} \right| = \prod_{V \in \mathbf{An}(\mathbf{X},G)} m_V(V, \mathbf{Parents}(V,G))$$

(This is because for an acyclic graph the transformation matrix can be arranged so that it is lower triangular.) Each term $|\partial \varepsilon_V / \partial V|$ is some non-negative function m_V of V and its parents, because ε_V is a function of V and its parents. Hence by lemma 1, if **X** and **Y** are d-separated given **Z**, then **X** and **Y** are independent given **Z**.

Suppose that **X** and **Y** are not d-separated given **Z**. Then by Theorem 2, there is a linear SEM in which **X** and **Y** are not independent given **Z**. Since a linear SEM is a special case of an SEM, there is an SEM in which **X** and **Y** are not independent given **Z**. ∴

Lemma 5: In an SEM with directed graph G with vertices **V**, if **X** is an ancestral set for G, then each minimally sufficient product of terms occurring in $J_{\mathbf{X} \to \mathbf{Err}(\mathbf{X})}$ that is non-zero is either a loop in G(**X**), or $\partial \varepsilon_V / \partial V$ for V in **X**.

Proof. Each term in $J_{\mathbf{X} \to \mathbf{Err}(\mathbf{X})}$ is a product of partial derivatives in the partial derivative matrix, one from each row, and one from each column, times a variable that is either equal to 1 or -1. Hence each variable in **X** appears exactly once in the numerator of some partial derivative in the term, and exactly once in the denominator of some partial derivative in the term. If $\partial \varepsilon_V / \partial V$ occurs in term T, it is minimally sufficient.

Suppose then that some minimally sufficient product of partial derivatives S occurring in term T is not equal to $\partial \varepsilon_V / \partial V$ for any V in **X**. Then S does not contain $\partial \varepsilon_V / \partial V$ for V in **X**, because otherwise it would not be minimally sufficient. Hence each partial derivative in S is of the form $\partial \varepsilon_V / \partial Y$ where $V \neq Y$. Such a term is non-zero only if there is an edge from Y to V in G. Because V and Y are both in ancestral set **X**, if there is an edge from Y to V in G, then there is an edge from Y to V in G(**X**). Since all of the occurrences of the variables in S are in S, each variable occurs once in the numerator and once in the denominator of a partial derivative in S; so in G(**X**) there is a path in which all of the variables in S occur once at the head of an edge and once at the tail. It follows that there is a cycle in G(**X**) that corresponds to the product of partial derivatives in S. ∴

Let the entry in the I[th] row and the/[h] column of matrix A be denoted *Ay.* Let

$$\varepsilon_{abc...n} \equiv \begin{cases} +1 & \text{if } (abc.n) \text{ is an even permutation of } (1,2,3...TV) \\ -1 & \text{if } (abc.n) \text{ is an odd permutation of } (l,2,3...A0 \\ 0 & \text{otherwise} \end{cases}$$

Then if *A* is an *N* x *N* matrix,

det(A) = *£abc...n^A\a^A2b-^ANn*

where each of the variables in the subscript of *e* range from 1 to *N*, and are summed over. If *(abc.n)* is some permutation of (1, 2, 3, ..., *N)*, then it follows that *Sabc.jn* = (-1)^ whre *P* is the number of transpositions of subscripts neeed to produce the order (1, 2, 3, ..., TV). See Byron and Fuller (1969).

Corresponding to each edge in *G* is an entry in the matrix of partial derivatives relating the error terms to the non-error terms. A **cycleset** in a directed graph G is a set of non-intersecting cycles. Hence, corresponding to a cycleset C in a graph G is a set of entries in the matrix of partial derivatives, one for each edge in the cycleset. For example, if a cycleset C contains a cycle *C* = «Xi,X4>,<X4>Xi», and C2 = *«X2Xs>,<XsXl»* then the corresponding matrix entries are A14, A41, A25, and A52. Order these matrix entries by row number. In this example, rearranging the matrix entries in order of rows yields <Ai4, A25, A41, A52X If the corresponding sequence of column numbers for cycleset C is *{abc.}*, let *P(C)* be the number of transpositions of subscipts needed to rearrange the column numbers in increasing order. In this example, the corresponding sequence of column numbers is <4, 5, 1, 2>. To rearrange these numbers into the order <1, 2, 4, 5> requires two transpositions. Hence *P(C)* = 2.

For convenience, if an edge *<W,Y>* occurs in a cycle C, write *<WJ> e* C, even though strictly speaking, edges are not members of cycles. Let Cydeset(G) be the set of all cyclesets in G. Let **Vertices(C)** be the set of vertices occurring in a set of cycles C.

Lemma 6: In an SEM with directed graph G with vertices V, if X is an ancestral set in G then

$$J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})} = \sum_{\mathbf{C} \in \mathbf{Cycleset}(G(\mathbf{X}))} -1^{P(\mathbf{C})} \times \left(\prod_{<W,Y> \in \mathbf{C}} \frac{\partial \varepsilon_Y}{\partial W} \right) \left(\prod_{V \in \mathbf{X} \setminus \mathbf{Vertices}(\mathbf{C})} \frac{\partial \varepsilon_V}{\partial V} \right)$$

Proof. For each **C** that is a set of cycles in $G(\mathbf{X})$ that do not intersect, let

$$g(\mathbf{C}) = -1^{P(\mathbf{C})} \times \left(\prod_{<W,Y> \in \mathbf{C}} \frac{\partial \varepsilon_Y}{\partial W} \right) \left(\prod_{V \in \mathbf{X} \setminus \mathbf{Vertices}(\mathbf{C})} \frac{\partial \varepsilon_V}{\partial V} \right)$$

I will show that for each cycleset **C** in $G(\mathbf{X})$ that $g(\mathbf{C})$ is a term in $J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})}$, every non-zero term in $J_{\mathbf{V}\text{->}\mathbf{Err}(\mathbf{V})}$ is equal to $g(\mathbf{C})$ for some cycleset **C** in $G(\mathbf{X})$, and if $\mathbf{C}_1$ and $\mathbf{C}_2$ are distinct cyclesets then $g(\mathbf{C}_1) \neq g(\mathbf{C}_2)$.

For each **C**, a variable occurs once in the denominator of a partial derivative in $g(\mathbf{C})$, and once in the numerator of partial derivative in $g(\mathbf{C})$. Hence one partial derivative from each row and each column of the transformation matrix occurs in $g(\mathbf{C})$. But every product of partial derivatives which consists of one partial derivative from each column and each row of the transformation matrix is a term in $J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})}$ (because $J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})}$ is the determinant of the transformation matrix). Hence $g(\mathbf{C})$ is a term in $J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})}$.

Let $\mathbf{C}_1$ be a set of cycles such that no pair of cycles in $\mathbf{C}_1$ intersect, and similarly for $\mathbf{C}_2$. Suppose that $\mathbf{C}_1 \neq \mathbf{C}_2$; then $g(\mathbf{C}_1) \neq g(\mathbf{C}_2)$ unless there is some way to rearrange the edges in $\mathbf{C}_1$ into the cycles in $\mathbf{C}_2$. But because no pair of cycles in $\mathbf{C}_1$ intersect, each vertex that appears in $\mathbf{C}_1$ occurs in exactly two edges, once as the head, and once as the tail. Hence the edges in $\mathbf{C}_1$ cannot be rearranged into the cycles in $\mathbf{C}_2$, and $g(\mathbf{C}_1) \neq g(\mathbf{C}_2)$.

By lemma 5, each minimally sufficient product of terms occurring in T of $J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})}$ is either a loop or $\partial \varepsilon_V / \partial V$ for V in **X**. By definition, the variables in distinct minimally sufficient product of terms do not overlap. Hence T consists of a product of non-intersecting minimally sufficient products of terms. Hence, for every non-zero term T in $J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})}$ there is a cycleset **C** such that $T = g(\mathbf{C})$.

For a given term in the sum, P is equal to the number of transpositions needed to arrange the column numbers in increasing order. The column numbers corresponding to vertices not in **Vertices**(**C**) do not need any rearrangement. Hence P for a given term is equal to $P(\mathbf{C})$. $\therefore$

Let **Cyclegroup**(G) be the set of all cyclegroups in directed graph G. If **C** is a cyclegroup in G, let **Cycleset(C)** be the set of all cyclesets included in **C**. Let **Cycles**(G) be the set of all vertices that occur on cycles in G.

Lemma 7: In an SEM with directed graph G, if **X** is an ancestral set for G, then

$$J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})} =$$

$$\left(\prod_{V \notin \mathbf{Cycles}(G(\mathbf{X}))} \frac{\partial \varepsilon_V}{\partial V} \right) \times$$

$$\left(\prod_{\mathbf{C} \in \mathbf{Cyclegroup}(G(\mathbf{X}))} \left(\sum_{\mathbf{D} \in \mathbf{Cycleset}(\mathbf{C})} \left(-1^{P(\mathbf{D})} \times \left(\prod_{V \in \mathbf{Vertices}(\mathbf{C} \setminus \mathbf{D})} \frac{\partial \varepsilon_V}{\partial V} \right) \left(\prod_{<W,Y> \in \mathbf{D}} \frac{\partial \varepsilon_Y}{\partial W} \right) \right) \right) \right)$$

Proof. By lemma 6,

$$J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})} = \sum_{\mathbf{C} \in \mathbf{Cycleset}(G(\mathbf{X}))} -1^{P(\mathbf{C})} \times \left(\prod_{<W,Y> \in \mathbf{C}} \frac{\partial \varepsilon_Y}{\partial W} \right) \left(\prod_{V \in \mathbf{X} \setminus \mathbf{Vertices}(\mathbf{C})} \frac{\partial \varepsilon_V}{\partial V} \right)$$

If V is not in a cycle in $G(\mathbf{X})$ then it is not in any cycleset. Hence, by lemma 5, every occurrence of V in each non-zero term in $J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})}$ is of the form $\partial \varepsilon_V / \partial V$. Hence it is possible to factor

$$\left(\prod_{V \notin \mathbf{Cycles}(G(\mathbf{X}))} \frac{\partial \varepsilon_V}{\partial V} \right)$$

from each non-zero term in the previous equation. This leads to

$$J_{\mathbf{X}\text{->}\mathbf{Err}(\mathbf{X})} =$$

$$\left(\prod_{V \notin \mathbf{Cycles}(G(\mathbf{X}))} \frac{\partial \varepsilon_V}{\partial V} \right) \times$$

$$\left(\sum_{\mathbf{C} \in \mathbf{Cycleset}(G(\mathbf{X}))} -1^{P(\mathbf{C})} \times \left(\prod_{<W,Y> \in \mathbf{C}} \frac{\partial \varepsilon_Y}{\partial W} \right) \left(\prod_{V \in \mathbf{Cycles}(G(\mathbf{X})) \setminus \mathbf{Vertices}(\mathbf{C})} \frac{\partial \varepsilon_V}{\partial V} \right) \right)$$

The set of cyclegroups in G partitions the set of cycles in G. Hence each cycleset in G can be partitioned into a set of cyclesets, where each cycleset contains only cycles from the same cyclegroup. In addition, suppose that C is a set of cyclesets, where each cycleset in C contains cycles from only one cyclegroup, and each pair of cyclesets in C contains cycles from different cyclegroups. Then the union of any two cyclesets in C is also a cycleset. Hence,

$$\sum_{C \in Cycleset(G(X))} -1^{P(C)} \times \left(\prod_{\langle W,Y\rangle \in C} \frac{\partial \varepsilon_Y}{\partial W}\right)\left(\prod_{V \in Cycles(G(X)) \setminus Vertices(C)} \frac{\partial \varepsilon_V}{\partial V}\right) =$$

$$\prod_{C \in Cyclegroup(G(X))} \left(\sum_{D \in Cycleset(C)} -1^{P(D)} \times \left(\prod_{V \in Vertices(C \setminus D)} \frac{\partial \varepsilon_V}{\partial V}\right)\left(\prod_{\langle W,Y\rangle \in D} \frac{\partial \varepsilon_Y}{\partial W}\right)\right)$$

∴

The following example illustrates an application of lemma 7. Consider the directed graph shown in figure 7. There are two cyclegroups consisting of $\{C_1, C_2, C_3\}$ and $\{C_4, C_5\}$. The set of all cyclesets included in the first cyclegroup is $\{\emptyset, \{C_1\}, \{C_2\}, \{C_3\}, \{C_1, C_3\}\}$, and the set of all cyclesets included in the second cyclegroup is $\{\emptyset, \{C_4\}, \{C_5\}\}$. The matrix of the partial derivatives is shown below.

$$\begin{vmatrix}
\frac{\partial\varepsilon_{X_1}}{\partial X_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{\partial\varepsilon_{X_2}}{\partial X_1} & \frac{\partial\varepsilon_{X_2}}{\partial X_2} & \frac{\partial\varepsilon_{X_2}}{\partial X_3} & \frac{\partial\varepsilon_{X_2}}{\partial X_4} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{\partial\varepsilon_{X_3}}{\partial X_2} & \frac{\partial\varepsilon_{X_3}}{\partial X_3} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{\partial\varepsilon_{X_4}}{\partial X_2} & 0 & \frac{\partial\varepsilon_{X_4}}{\partial X_4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{\partial\varepsilon_{X_5}}{\partial X_4} & \frac{\partial\varepsilon_{X_5}}{\partial X_5} & 0 & 0 & 0 & 0 \\
0 & \frac{\partial\varepsilon_{X_6}}{\partial X_2} & 0 & 0 & 0 & \frac{\partial\varepsilon_{X_6}}{\partial X_6} & \frac{\partial\varepsilon_{X_6}}{\partial X_7} & \frac{\partial\varepsilon_{X_6}}{\partial X_8} & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{\partial\varepsilon_{X_7}}{\partial X_6} & \frac{\partial\varepsilon_{X_7}}{\partial X_7} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{\partial\varepsilon_{X_8}}{\partial X_6} & 0 & \frac{\partial\varepsilon_{X_8}}{\partial X_8} & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{\partial\varepsilon_{X_9}}{\partial X_6} & 0 & 0 & \frac{\partial\varepsilon_{X_9}}{\partial X_9}
\end{vmatrix}$$

Applying lemma 7, the Jacobean can be factored in the following way:

$$\frac{\partial\varepsilon_{X_1}}{X_1} \times \frac{\partial\varepsilon_{X_9}}{X_9} \times$$

$$\left[(-1)^0 \times \frac{\partial\varepsilon_{X_2}}{X_2} \times \frac{\partial\varepsilon_{X_3}}{X_3} \times \frac{\partial\varepsilon_{X_4}}{X_4} \times \frac{\partial\varepsilon_{X_5}}{X_5} + (-1)^1 \times \frac{\partial\varepsilon_{X_2}}{X_3} \times \frac{\partial\varepsilon_{X_3}}{X_2} \times \frac{\partial\varepsilon_{X_4}}{X_4} \times \frac{\partial\varepsilon_{X_5}}{X_5} + \right.$$

$$(-1)^1 \times \frac{\partial\varepsilon_{X_2}}{X_4} \times \frac{\partial\varepsilon_{X_4}}{X_2} \times \frac{\partial\varepsilon_{X_3}}{X_3} \times \frac{\partial\varepsilon_{X_5}}{X_5} + (-1)^1 \times \frac{\partial\varepsilon_{X_4}}{X_5} \times \frac{\partial\varepsilon_{X_5}}{X_4} \times \frac{\partial\varepsilon_{X_2}}{X_2} \times \frac{\partial\varepsilon_{X_3}}{X_3} +$$

$$\left.(-1)^2 \times \frac{\partial\varepsilon_{X_2}}{X_3} \times \frac{\partial\varepsilon_{X_3}}{X_2} \times \frac{\partial\varepsilon_{X_4}}{X_5} \times \frac{\partial\varepsilon_{X_5}}{X_4}\right] \times$$

$$\left[(-1)^0 \times \frac{\partial\varepsilon_{X_6}}{X_6} \times \frac{\partial\varepsilon_{X_7}}{X_7} \times \frac{\partial\varepsilon_{X_8}}{X_8} + (-1)^1 \times \frac{\partial\varepsilon_{X_6}}{X_7} \times \frac{\partial\varepsilon_{X_7}}{X_6} \times \frac{\partial\varepsilon_{X_8}}{X_8} + (-1)^1 \times \frac{\partial\varepsilon_{X_6}}{X_8} \times \frac{\partial\varepsilon_{X_8}}{X_6} \times \frac{\partial\varepsilon_{X_7}}{X_7}\right]$$

Lemma 8: For an SEM with directed graph *G* with vertices V, if X is an ancestral set in *G* then

$$lv(\mathbf{X}) =$$

$$\left(\prod_{V \notin \mathbf{Cycles}(G(\mathbf{X}))} g_V(V,\mathbf{Parents}(V,G(\mathbf{X})))\right) \times \left(\prod_{{}^\wedge \mathrm{CeCyclegroup(G(X))}} g_C(\mathbf{C},\mathbf{Parents}(\mathbf{C},G(\mathbf{X})))\right)$$

where each *gy* and *gQ* is a non-negative function.

Proof. The transformed density function of **Err(X)** is equal to

(1) $$\left(\prod_{X \in \mathbf{X}} f_{\mathbf{Err}}(h_X(X,\mathbf{Parents}(X,G(\mathbf{X}))))\right) \Big|_{\mathrm{Err(X)}\_\to x}$$

where $ex = hx(X,Parents(X,G(X)))$**. By lemma 7,**

(2) $$J_{\mathbf{X} \to \mathbf{Err}(\mathbf{X})} =$$

$$\left(\mathbf{n}_{[V \in \mathrm{Cycles(G(X))}} \frac{d\pounds y}{1V}\right) \times$$

$$\left(\mathbf{n}_{\mathrm{CeCyclegroup(G(X))}}\left(\sum_{\mathrm{D\ 6\ Cycleset(C)}}\left(-1^{P(\mathbf{D})} \times \left(\prod_{V \text{ e Vertices}(C \backslash D)} \frac{\partial \varepsilon_V}{\partial V}\right)\left(\prod^{11}_{<W,Y> \mathrm{GD}} \frac{\partial \varepsilon_Y}{\partial W}\right)\right)\right)\right)$$

Each term in

$$\left(\mathbf{n}_{V \notin \mathbf{Cycles}(G(\mathbf{X}))} \frac{de_v}{dV}\right)$$

is a function of *V* and ParentsO^GCX)). Each term in

$$\mathbf{n}_{\mathrm{CGCydegroup(G(X))}}\left(\mathbf{X}_{\mathrm{DeCydeset(C)}}\left(-1^{P(\mathbf{D})} \times \left(\mathbf{n}_{V \in \mathrm{Vertices}(C \backslash D)} \frac{de_v}{\sim dV}\right)\left(\mathbf{n}_{<W,Y> \in \mathbf{D}^t} \frac{deY\_)}{dw)}\right)\right)_{yy}\right)$$

contains only error terms associated with variables in C, and hence is a function of C and Parents(C,G(X)). Hence, there exist non-negative functions *my* such that

(3)
$$\left|J_{\mathbf{X}\to\mathbf{Err(X)}}\right| =$$

$$\left(\prod_{V \notin \mathbf{Cycles}(G(\mathbf{X}))} m_V(V,\mathbf{Parents}(V,G(\mathbf{X}))\right) \times \left(\prod_{C \in \mathbf{Cyclegroup}(G(\mathbf{X}))} m_{\mathrm{C}}(\mathrm{C},\mathbf{Parents}(\mathrm{C},G(\mathbf{X})))\right)$$

Because $J_{\mathbf{Err(X)}\to\mathbf{X}} = 1/J_{\mathbf{X}\to\mathbf{Err(X)}}$, $J_{\mathbf{Err(X)}\to\mathbf{X}}$ can also be factored as in 3. Combining this with (1), there exist functions non-negative functions g_V and g_{C} such that

$$f_{\mathbf{V}}(\mathbf{X}) = \left(\prod_{V \notin \mathbf{Cycles}(G(\mathbf{X}))} g_V(V,\mathbf{Parents}(V,G(\mathbf{X}))\right) \times \left(\prod_{\mathrm{C} \in \mathbf{Cyclegroup}(G(\mathbf{X}))} g_{\mathrm{C}}(\mathrm{C},\mathbf{Parents}(\mathrm{C},G(\mathbf{X})))\right)$$

$\therefore$

Theorem 5: If G is a directed graph (cyclic or acyclic), collapsed graph G' contains disjoint sets of variables **X**, **Y** and **Z**, and **X** is d-separated from **Y** given **Z** in G' then G pseudo-indeterministically entails that **X** is independent of **Y** given **Z**.
Proof. By lemma 8 there exist non-negative functions g_V and g_{C} such that

$$f_{\mathbf{V}}(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}, G)) =$$
$$\left(\prod_{V \notin \mathbf{Cycles}(G(\mathbf{An}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z},G)))} g_V(V,\mathbf{Parents}(V,G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G))\right) \times$$
$$\left(\prod_{\mathrm{C} \in \mathbf{Cyclegroup}(G(\mathbf{An}(\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z},G)))} g_{\mathrm{C}}(\mathrm{C},\mathbf{Parents}(\mathrm{C},G(\mathbf{An}(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z},G))))\right)$$

This is a factorization according to the collapsed graph G', and hence by lemma 1, for three disjoint sets of variables **X**, **Y** and **Z**, if **X** and **Y** are d-separated given **Z** in G', then **X** and **Y** are independent given **Z**. $\therefore$

Theorem 6: If $P_{Mix}(\mathrm{V})$ is a mixture of probability measures, each of which factors according to directed graph G_i over **V**, $P_{Mix}(\mathrm{V})$ satisfies the global directed Markov property for G_{Mix}.
Proof. The density of the mixed probability measure can be represented in the following form. Introduce a variable T which takes on the value i in the i^{th} subpopulation. Denote the density function of the mixed probability measure by f_{Mix}. Set $g_{Mix,T}(T) = f_{Mix}(T)$, the

density of the individual subpopulations in the mixture. If *T* is not a parent of *V* in *GMU* set *gMix,v(V,Parents(V,GMix))* = gi,vO^Parents(V,G/)) (which in this case is the same for **all *i)*. If *T* is in *Pzrents(V$_y$G$_M$ix)* set *g$_M$ix,v(y,Parents(V,G$_M$ix))* = £/y(K,Parents(^G;))** for the value *T* = /. (Note that in the latter case the set of variables that are arguments to the function *gMix,V* ^{m}&y be a superset of the set of variables that are arguments to the function *giy$_y$* but the value of *gMixy* for *T* = / is determined by the subset of its arguments that are also arguments to *giy.)*

Suppose first that *An(X$_y$GMix)* does not contain *T.* Then each G/ is the same, each/; is the same, and

$$f_{Mix}(An(X\backslash G_{Mix}) = /KAnCX, <\%)) =$$

$$\prod_{XeAn(X \&))} g_{i,X}(X,\mathrm{Parents}(X,G_i)) = \prod_{XeAn(X,G_{Mix}))} g_{Mix,X}(X,\mathrm{Parents}(X,G_{Mix}))$$

Suppose next that *T* is in AII(X,GMU:)- Consider)3i«x(An(X,GAfix)) for the value *T* = /. Note that for *T* = *i$_y$fi(An(X,GMixMT})* is equal to/M/x(An(X,GAfLcA{*T}\T=i).* Assuming ***T = i*,**

$$f_{Mix}(An(X,G_{Mix})) = f_{Mix}(An(X,G_{Mix})\backslash \{T\}\backslash T = i) \times f_{Mix}(T = i)^{\cdot} =$$

$$/_{/}(\mathrm{An}(X, G_{Mix}) \backslash \{T\}) \times f_{Mix}\{T = i)$$

If there exists a set R such that An(R,G/) = *An(X$_y$GMix)\{T)* then by hypothesis *fi(An(X$_y$GMix)\{T})* can be factored into a product of non-negative functions of members of An(R,G/) and their parents. I will show that such a set R exists. Let R = *An(X$_y$GMix)\{T}.* Then An(X,GMa)M^} £ An(R,G/) by definition of the ancestor relation. G/ is a subgraph of *GMIX* that does not contain 7\ so every ancestor of a member of R in *G[* is an ancestor of a member of X in *GMIX-* Hence An(R,G;)_e *An(X,GMix)\{T}.* It follows that An(R,G/) = An(X,GMu:)M7'}, and//(An(X,GMa)M^}) can be factored in the following way, where each *gMixJC* is a non-negative function:

$$f_i(\mathbf{An}(\mathbf{X},G_{Mix}) \setminus \{T\}) \times f_{Mix}(T = i) =$$

$$\prod_{X \in \mathbf{An}(\mathbf{X},G_{Mix})) \setminus \{T\}} g_{i,X}(X, \mathbf{Parents}(X,G_i)) \times g_{Mix,T}(T = i) =$$

$$\prod_{X \in \mathbf{An}(\mathbf{X},G_{Mix}))} g_{Mix,X}(X, \mathbf{Parents}(X,G_{Mix}))$$

Hence by lemma 1, $f_{Mix}(\mathbf{V},G_{Mix})$ is represented by G_{Mix}. $\therefore$

References

Bollen, K. (1989). Structural Equations with Latent Variables. Wiley, NY.

Byron, F., and Fuller, R. (1969) Mathematics of Classical and Quantum Physics. Addison Wesley, MA.

Cooper, G., and Herskovits, E. (1992) "A Bayesian Method for the Induction of Probabilistic Networks from Data." *Machine Learning*, **9**, 309-348.

Frydenberg, M. (1990). "The Chain Graph Markov Property." *Scand. J. Statist*, **17**, 333-353.

Geiger, D., and Pearl, J. (1988). "Logical and Algorithmic Properties of Conditional Independence." Technical Report R-97, Cognitive Systems Laboratory, University of California, Los Angeles.

Goldberger, A., Duncan, O. (eds.) (1973). Structural Equation Models in the Social Sciences. Seminar Press, NY.

Haavelmo, T. (1943). "The Statistical Implications of a System of Simultaneous Equations." *Econometrica*, **11**, 1-12.

Kiiveri, H. and Speed, T. (1982). "Structural Analysis of Multivariate Data: A Review." *Sociological Methodology*, Leinhardt, S. (ed.). Jossey-Bass, San Francisco.

Kiiveri, H., Speed, T., and Carlin, J. (1984). "Recursive Causal Models." *Journal of the Australian Mathematical Society*, **36**, 30-52.

Lauritzen, S., Dawid, A., Larsen, B., Leimer, H. (1990). "Independence Properties of Directed Markov Fields." *Networks*, **20**, 491-505.

Mason, S., (1953). "Feedback Theory-Some Properties of Signal Flow Graphs." *Proceedings of the IRE*, **41**.

Mason, S., (1956). "Feedback Theory-Further Properties of Signal Flow Graphs." *Proceedings of the IRE*, **44**.

Pearl, J. (1986). "Fusion, Propagation, and Structuring in Belief Networks." *Artificial Intelligence* **29**, 241-88.

Pearl, J., (1988). Probabilistic Reasoning in Intelligent Systems, Morgan Kaufman: San Mateo, CA.

Pearl, J. and Verma, T. (1991). "A Theory of Inferred Causation." Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference, Morgan Kaufmann, San Mateo, CA.

Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R. (1993) "Bayesian Analysis in Expert Systems.", *Statisical Science*, **8**, 219-246.

Spirtes, P. and Glymour, C. (1990). "Causal Structure Among Measured Variables Preserved with Unmeasured Variables." Technical Report CMU-LCL-90-5, Laboratory for Computational Linguistics, Carnegie Mellon University.

Spirtes, P., and Glymour, C., (1991). "An Algorithm for Fast Recovery of Sparse Causal Graphs." *Social Science Computer Review,* **9**, 62-72.

Spirtes, P., Glymour, C., and Scheines, R. (1993). Causation, Prediction, and Search, Springer-Verlag Lecture Notes in Statistics 81, NY.

Wermuth, N. (1980). "Linear Recursive Equations, Covariance Selection and Path Analysis." *Journal of the American Statistical Association*, **75**, 963-972.

Wermuth, N. and Lauritzen, S. (1983). "Graphical and Recursive Models for Contingency Tables." *Biometrika*, **72**, 537-552.

Wermuth, N. and Lauritzen, S. (1990). "On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models." *J. Roy. Statist. Soc. Ser. B*, **52**, 21-50.

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics, Wiley, NY.

Wright, S. (1934). "The Method of Path Coefficients." *Ann. Math. Stat.* **5**, 161-215.