Chain Graph Models and their Causal Interpretations*

Steffen L. Lauritzen† Aalborg University

Thomas S. Richardson University of Washington

February 2001

Abstract

Chain graphs are a natural generalization of directed acyclic graphs (DAGs) and undirected graphs. However, the apparent simplicity of chain graphs belies the subtlety of the conditional independence hypotheses that they represent. There are a number of simple and apparently plausible, but ultimately fallacious interpretations of chain graphs that are often invoked, implicitly or explicitly. These interpretations also lead to flawed methods for applying background knowledge to model selection. We present a valid interpretation by showing how the distribution corresponding to a chain graph may be generated as the equilibrium distribution of dynamic models with feedback. These dynamic interpretations lead to a simple theory of intervention, extending the theory developed for DAGs. Finally, we contrast chain graph models under this interpretation with simultaneous equation models which have traditionally been used to model feedback in econometrics.

KEYWORDS: Causal model; chain graph; feedback system; intervention theory; Gibbs sampler; simultaneous equation model.

1 Introduction

The use of directed acyclic graphs (DAGs) to simultaneously represent causal hypotheses and to encode independence and conditional independence constraints associated with those hypotheses may be traced back to the pioneering work of Sewall Wright (1921). More recently, DAGs have proved

*This is Research Report R-01-2003, Department of Mathematical Sciences, Aalborg University.

†Address for correspondence: Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9200 Aalborg, Denmark.

fruitful in the construction of expert systems, in the development of efficient updating algorithms (Pearl 1988; Lauritzen and Spiegelhalter 1988), and reasoning about causal relations (Pearl 1988, 1995, 2000; Lauritzen 2001; Spirtes *et al.* 1993).

Graphical models based on undirected graphs, also called Markov random fields, have been used in spatial statistics to analyse data from field trials, image processing, and a host of other applications (Hammersley and Clifford 1971; Besag 1974b; Speed 1979; Darroch *et al.* 1980).

Chain graphs, which admit both directed and undirected edges, but no partially directed cycles, were introduced as a natural generalization of both undirected graphs and acyclic directed graphs (Lauritzen and Wermuth 1989). One of the original motivations for introducing chain graphs was that the inclusion of undirected edges allowed the modelling of "simultaneous responses" (Frydenberg 1990), "symmetric associations" (Lauritzen and Wermuth 1989) or simply "associative relations", as distinct from causal relations (Andersson *et al.* 1996), represented by directed edges.

Chain graph models are beginning to be used increasingly in applied contexts, see e.g. Mohamed *et al.* (1998). A central theme of this paper is that the apparent simplicity of chain graphs as an extension of DAGs and undirected graphs belies the subtlety of the hypotheses that they represent. In particular, there are a number of simple and apparently plausible, but ultimately fallacious and misleading, interpretations of chain graphs that are often invoked implicitly or explicitly as a justification for their application. In Section 5 we describe and discuss such interpretations.

We next present valid interpretations, by showing how the distribution corresponding to a chain graph may be generated as equilibrium distributions of dynamic models with feedback over time. Here again we will see that things are not quite as straightforward as they may at first appear.

This dynamic interpretation leads to a simple theory of intervention, extending the theory developed for DAGs. Finally, we contrast chain graph models with simultaneous equation models which have traditionally been used to model feedback in econometrics.

2 Basic Graphical Concepts and Notation

We consider graphs containing both directed ($\rightarrow$) and undirected edges ($\,—\,$). A *partially directed cycle* in a graph $\mathcal{G}$ is a sequence of n distinct vertices $v_1, \ldots v_n$, $(n \geq 3)$, and $v_{n+1} \equiv v_1$, such that (a) $\forall i$ $(1 \leq i \leq n)$ either $v_i — v_{i+1}$ or $v_i \leftarrow v_{i+1}$, and (b) $\exists j$ $(1 \leq j \leq n)$ such that $v_j \leftarrow v_{j+1}$. A *chain graph,* (CG), is a graph in which there are no partially directed cycles. A chain graph in which there are no undirected edges is a *directed acyclic graph,* most often abbreviated as DAG.

The *chain components* $\mathcal{T}$ of a chain graph are the undirected graphs obtained by removing all directed edges from the chain graph. A *minimal complex* in a chain graph is an induced subgraph of the form

$$a \rightarrow v_1 — \quad \cdots \qquad\qquad \cdots — v_r \leftarrow b.$$

3 Graphical Models

A *graphical model* is formally a set of distributions, satisfying a set of conditional independence relations encoded by a graph. This encoding is known as the *Markov property* associated with the type of graph. This article is concerned with the chain graph Markov property defined in Lauritzen and Wermuth (1984, 1989) and Frydenberg (1990). There have been a number of alternative suggestions for associating a Markov property with a chain graph (Cox and Wermuth 1993; Andersson *et al.* 1996), which generally are not equivalent to the above and which are not discussed in detail in the present paper.

Below we give the factorization versions of the Markov properties for DAGs and for chain graphs. For further details, the reader is referred to Lauritzen (1996).

3.1 Basic factorizations

A distribution P satisfying the Markov property associated with a DAG is most easily described through the factorization of its joint density f with respect to a product measure given by

$$f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}). \tag{1}$$

In terms of factorization, the chain graph Markov property manifests itself through an outer factorization

$$f(x) = \prod_{\tau \in \mathcal{T}} f\left(x_\tau \mid x_{\mathrm{pa}(\tau)}\right), \tag{2}$$

where each factor further factorizes according to the graph as

$$f\left(x_\tau \mid x_{\mathrm{pa}(\tau)}\right) = Z^{-1}\left(x_{\mathrm{pa}(\tau)}\right) \prod_{A \in \mathcal{A}(\tau)} \phi_A(x_A), \tag{3}$$

where $\mathcal{A}(\tau)$ are the complete sets in the subgraph $\mathcal{G}_{\tau \cup \mathrm{pa}(\tau)}$ and

$$Z\left(x_{\mathrm{pa}(\tau)}\right) = \sum_{x_\tau} \prod_{A \in \mathcal{A}(\tau)} \phi_A(x_A).$$

Note that the outer factorization (2) may be viewed as a directed acyclic graph with vertices representing the multivariate random variables X_τ for $\tau \in \mathcal{T}$. Andersson *et al.* (1996) refer to this as the "DAG of boxes" associated with a chain graph, but 'DAG of chain components' would be more precise, as boxes typically are used to indicate a coarser partitioning of the variables than specified with chain components (Wermuth and Lauritzen 1990).

3.2 The global Markov property and Markov equivalence

The *global Markov property* associated with a DAG $\mathcal{D}$ or a chain graph $\mathcal{K}$ identifies the full set of conditional independence relations that follow as consequences of the factorizations above.

In general, different graphs can imply the same conditional independence relations. More precisely, if for given state spaces we let $M(\mathcal{G})$ denote the set of distributions obeying the conditional independence relations associated with a graph $\mathcal{G}$, two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be *Markov equivalent* if $M(\mathcal{G}_1) = M(\mathcal{G}_2)$ for all such state spaces. Frydenberg (1990) gave the following necessary and sufficient condition for Markov equivalence of two chain graphs:

Proposition 1 *Two chain graphs $\mathcal{K}_1$ and $\mathcal{K}_2$ are Markov equivalent if and only if they have the same adjacencies and the same minimal complexes.*

The similar result for DAGs was obtained by Verma and Pearl (1990).

4 Causal Interpretation of DAG Models

This section gives a brief description of the now rather standard causal interpretations associated with a DAG given by Spirtes *et al.* (1993) and Pearl (1993, 1995), largely following Lauritzen (2001). The interpretations are both concerned with their *data generating processes* and associated calculation of average *effects of interventions.*

4.1 Conditioning by observation or intervention

We initially emphasize the distinction between different types of conditioning operations, each of which modify a given probability distribution in response to information obtained. Conditional densities are sometimes calculated as

$$f(y \mid x) = f(y \mid X = x) = f(y, x)/f(x).$$

We refer to this type of conditioning as *conditioning by observation* or *conventional conditioning.*

This is typically not the way the distribution of Y should be modified if we intervene externally and force the value of X to be equal to x. We refer to this other type of modification as *conditioning by intervention* or *conditioning by action.* To make the distinction clear we use different symbols for this conditioning, as indicated below

$$f(y \,||\, x) = f(y \mid X \leftarrow x).$$

Generally, the two quantities will be different

$$f(y \,||\, x) \neq f(y \mid x)$$

and the quantity on the left-hand side cannot be calculated from the density alone, without additional assumptions.

Below we will give a precise causal interpretation of a directed acyclic graph. This will imply that in the graph below to the left

$$X \longrightarrow Y \qquad\qquad X \longleftarrow Y$$

we will have that $f(y \,||\, x) = f(y \mid x)$ and $f(x \,||\, y) = f(x)$, whereas these relations are reversed in the graph to the right, i.e. there it holds that $f(y \,||\, x) = f(y)$ and $f(x \,||\, y) = f(x \mid y)$.

4.2 Data generating process for DAG models

A *data generating process* for a DAG model is a system of assignments

$$X_v \leftarrow g_v(X_{\mathrm{pa}(v)}, U_v), v \in V, \tag{4}$$

where the assignments have to be carried out sequentially in a well-ordering of the directed acyclic graph $\mathcal{D}$, or partly in parallel, so that at all times, when X_v is about to be assigned a value, all variables in $\mathrm{pa}(v)$ have already been assigned a value. The variables $U_v, v \in V$ are assumed to be independent.

This assignment system can be seen as a generic structural equation model (Bollen 1989) as invented in the context of genetics (Wright 1921), and exploited in economics (Haavelmo 1943; Wold 1954) and social sciences (Goldberger 1972). The term 'equation' is really misplaced, and 'structural assignment model' would have been much more appropriate. Much controversy in the literature is due to treating the assignment systems as equation systems, 'solving' them and uncritically moving variables between the right-hand side and the left-hand side of (4). In particular, this matters when interventions are considered. See for example Pearl (1998) and Spirtes *et al.* (1998) for further discussion.

Structural equation models were also used as the main justification and motivation for studying causal Markov models in Kiiveri *et al.* (1984) and Kiiveri and Speed (1982).

It is appropriate to think of a data generating process as a 'computer program', writing (4) as

For $i = 1, \ldots, p$;

$\epsilon \leftarrow$ runif;

$x_i \leftarrow h_i(x_{\mathrm{pa}(i)}, \epsilon)$;

Return x;

Here `runif` denotes a random variable which is uniformly distributed on the unit interval and h_i is chosen so that if E has this distribution then $h_i(x_{\mathrm{pa}(i)}, E)$ has the same distribution as $g_i(x_{\mathrm{pa}(i)}, U_i)$.

It is an important aspect of structural equation models that they also specify the way in which intervention is to be carried out. As is implicit in much literature and, for example, quite explicit in Strotz and Wold (1960), the effect of the intervention $X_a \leftarrow x_a^*$ on a variable with label a is simply that the corresponding line in (4) or the equivalent computer program is replaced with the assignment described by the intervention. We refer to this type of intervention as *intervention by replacement.*

4.3 Causal DAGs

When we say that a DAG $\mathcal{D}$ is *causal* for a probability distribution P, we imply that it holds for any $A \subseteq V$ that

$$f(x_{V\setminus A} \,||\, x_A) \quad = \prod_{v \in V\setminus A} f(x_v \mid x_{\mathrm{pa}(v)}) = \frac{f(x)}{\prod_{v\in A} f(x_v \mid x_{\mathrm{pa}(v)})}. \qquad (5)$$

Note that for $A = \emptyset$ this says that P is Markov with respect to $\mathcal{D}$.

We also use the expression that P is a *causal directed Markov field* with respect to $\mathcal{D}$ or say that P is *causally Markov* with respect to $\mathcal{D}$. Thus the causal Markov property gives a relation between different probability measures, each representing the probability law associated with a specific intervention.

We will refer to (5) as the *intervention formula* for DAGs. It appeared in various forms in Spirtes *et al.* (1993) and Pearl (1993). It is implicit in Robins (1986) and in other literature.

Intervention by replacement conforms well with the intervention formula (5) as stated formally in the theorem below, which is Theorem 2.20 of Lauritzen (2001).

Proposition 2 *Let $X = (X_v)_{v \in V}$ be determined by a structural assignment system corresponding to a given directed acyclic graph $\mathcal{D}$ and let P denote its distribution. If intervention is carried out by replacement, then P is causally Markov with respect to $\mathcal{D}$.*

Thus in the case of a DAG there is full harmony between the causal interpretations determined by data generating processes, intervention by replacement, and the causal Markov property associated with the DAG.

5 Rationale for Chain Graphs and their Misuse

The modern theory of graphical models, in which a graph is used to represent a set of distributions, with independence structure encoded by a graph was originally developed using undirected graphs (Darroch *et al.* 1980).

In early applications of undirected graphical models, see e.g. Edwards and Kreiner (1983), the hypotheses of interest were in some sense causal, studying relationships beween explanatory and response variables. It is clearly unnatural to try to represent a system of such relations, which are asymmetric, by an undirected graph in which all relations are symmetric.

This motivated the development of graphical models with directed edges, thereby extending the work of Sewall Wright on path diagrams, and the theory of recursive structural equation models in econometrics (Wold 1953).

A pair of variables x, y in a set V, may be said to be *directly associated* (relative to V), if there is no $Z \subseteq V \setminus \{x, y\}$ so that $x \perp\!\!\!\perp y \mid Z$. Typically, if x and y are directly associated then the vertices are joined by an edge in a graphical model representing this distribution. However, as every student learns, association does not imply causation. Consequently, if directed edges are used to denote causal relations then it appears overly restrictive

(i) (ii)

Figure 1: Two examples of chain graphs in which c and d are joint responses to a and b.

to consider graphs in which all edges are directed, since to do so would rule out the possibility of non-causal associations. This motivates the inclusion of undirected edges within the graphs.

However, there are many different reasons why we may not wish to put a directed edge between two directly associated variables x, y. For example:

(i) The association may have arisen due to the presence of:

- an unmeasured confounding variable;
- some artefact of the way the sample was selected;
- a feedback relationship.

(ii) We may believe that the association is causal but not know whether x causes y or vice-versa.

There is a simple qualitative difference between (i) and (ii): in situation (ii), additional knowledge might justify including an edge $x \rightarrow y$, whereas this is not so, with (i). In philosophical terms, reasons under (i) would be described as *ontological*, those under (ii) as *epistemological*.

Although the original papers on chain graphs are clear that directed edges are to be interpreted as (in some sense) causal, while undirected edges are to represent non-causal associations, in which variables are 'on an equal footing' this leaves room for ambiguity because, as we have seen, non-causal associations may arise in many different ways.

The chain graph CG_2 in Figure 1 (i) corresponds to the following factorization of the joint density (assuming that the relevant conditional densities exist):

$$f(a, b, c, d) = f(c, d \,|\, a, b) f(a) f(b).$$

In this sense the model treats c, d as on an equal footing, as it places no restriction on the form of the conditional density $f(c, d \,|\, a, b)$. However, when submodels are considered, special attention is required. A submodel

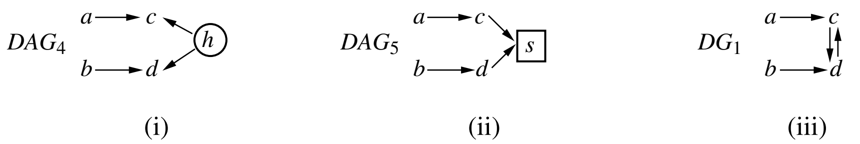


Figure 2: (i) &(ii) Generating processes in which c and d are 'on an equal footing', that do not give rise to the conditional independence model given by CG_3 under the standard Markov property. (iii) A directed cyclic graph, corresponding to a non-recursive linear structural equation model, again not Markov equivalent to CG_3.

such as CG_3 in Figure 1 (ii), restricts $f(c,d\,|\,a,b)$. Under the chain graph Markov property, CG_3 implies

$$a \perp\!\!\!\perp b \qquad a \perp\!\!\!\perp d\,|\,\{b,c\} \qquad b \perp\!\!\!\perp c\,|\,\{a,d\}$$

and, as we shall see, the undirected edge in this chain graph can not be interpreted in any of the ways listed above other than feedback.

For example, one might think that the chain graph structure displayed in CG_3 could be explained by one of the data-generating processes associated with the DAGs shown in Figure 2, (i) and (ii). In DAG_4 c and d share an unmeasured common parent; in the marginal distribution over the remaining variables

$$a \perp\!\!\!\perp \{b,d\}, \quad b \perp\!\!\!\perp \{a,c\} \quad \text{but} \quad a \not\!\perp\!\!\!\perp d\,|\,\{b,c\}, \quad b \not\!\perp\!\!\!\perp c\,|\,\{a,d\}$$

corresponding to an independence structure different from that of CG_3.

In DAG_5, c and d share a common child that has been conditioned on. In the conditional distribution of the remaining variables, given s:

$$a \perp\!\!\!\perp d\,|\,\{b,c\}, \quad b \perp\!\!\!\perp c\,|\,\{a,d\} \quad \text{but} \quad a \not\!\perp\!\!\!\perp b.$$

Consequently, neither of these generating processes explain CG_3 of Figure 1 (ii).

The directed cyclic graph in Figure 2 (iii) corresponds to a non-recursive linear structural equation model, see Section 7, Spirtes (1995) and Koster (1996) for further discussion of these models. The following independence relations hold in this model:

$$a \perp\!\!\!\perp b, \qquad a \perp\!\!\!\perp b\,|\,\{c,d\} \quad \text{but} \quad a \not\!\perp\!\!\!\perp d\,|\,\{b,c\}, \quad b \not\!\perp\!\!\!\perp c\,|\,\{a,d\},$$

which again does not correspond to CG_3.

In all three examples there is dependence between c and d, and these variables might be argued to be on an equal footing. Thus, CG_3 does not merely assert that c and d are on an equal footing, but a very particular kind of equal footing. This point was made by Cox and Wermuth (1993), who used it as a motivation for introducing alternative Markov properties for chain graphs.

5.1 Non-causal associations due to latent variables

One can strengthen the message in the examples above to say that there is no (finite) DAG model which, under marginalizing and conditioning, gives the set of conditional independence relations implied by CG_3. This was pointed out by Richardson (1998), who shows that all conditional independence structures which can be obtained by such marginalization and conditioning satisfy a property of *between separation* (Theorem 1 of loc. cit.), whereas CG_3 does not.

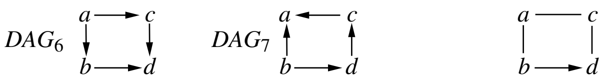
Although not using the terminology of chain graphs, Kiiveri *et al.* (1984) introduced the notion of a *recursive causal graph* as a chain graph where all chain components which were not singletons had no parents. Variables without parents were *exogenous* variables, i.e. variables that set the initial conditions for development of the remaining variables forming a recursive system determined by a DAG.

One can show (Richardson 2001) that such recursive causal graphs exactly correspond to the chain graphs obtainable from some DAG by marginalization and conditioning, as stated more accurately in the proposition below.

Proposition 3 *A chain graph $\mathcal{K}$ over the variables V represents the same set of conditional independence relations as derived from marginalizing over a set of variables L and conditioning on $X_S = x_S$ in a set of distributions represented by a directed acyclic graph $\mathcal{D}$ over $V \cup L \cup S$, if and only if it is Markov equivalent to a recursive causal graph.*

5.2 Chain graphs as unions of DAG models

Chain graph models are sometimes proposed as being appropriate in situations in which it is known that an edge is present, but the appropriate orientation of the edge is unknown. Such circumstances may for example arise during the construction of expert systems when a directed acyclic graph is elicited from an expert (Jensen 1996; Spiegelhalter *et al.* 1993).



(i) (ii)

Figure 3: (i) Two DAGs with the same sets of adjacencies. (ii) The graph formed from (i) by representing edges of different direction with undirected edges.

If $\mathcal{D}_1$ and $\mathcal{D}_2$ are two DAGs with the same set of adjacencies, but for some pair of vertices a, b, $a \leftarrow b$ in $\mathcal{D}_1$, but $a \rightarrow b$ in $\mathcal{D}_2$, then the graph $\mathcal{D}_{1\cup 2}$ obtained by representing common adjacencies of different directions with undirected edges may contain edges of both types. However, as exemplified in Figure 3, the graph produced in this way will only be a chain graph in quite special cases. Two such cases are:

(a) when $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent (Frydenberg 1990);

(b) when $\mathcal{D}_1$ and $\mathcal{D}_2$ possess the same adjacencies, but differ over the orientation of a single edge only.

However, even if the graph $\mathcal{D}_{1\cup 2}$ is a chain graph, this does not imply that the model determined by $\mathcal{D}_{1\cup 2}$ is equal to the union of the models determined by $\mathcal{D}_1$ and $\mathcal{D}_2$. In fact, if we let $M(\mathcal{G})$ denote the set of distributions obeying the Markov property associated with a graph $\mathcal{G}$ and assume that all state spaces have at least two elements, we have the following:

Proposition 4 *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two DAGs with the same adjacencies, such that $\mathcal{D}_{1\cup 2}$ is a chain graph. Then*

$$M(\mathcal{D}_{1\cup 2}) = M(\mathcal{D}_1) \cup M(\mathcal{D}_2)$$

if and only if $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent, i.e. when $M(\mathcal{D}_1) = M(\mathcal{D}_2)$.

Proof Frydenberg (1990) shows that if $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent then they are also Markov equivalent to $\mathcal{D}_{1\cup 2}$, proving one direction.

Conversely, if $\mathcal{D}_1$ and $\mathcal{D}_2$ are not Markov equivalent, but contain the same adjacencies, then it follows from Proposition 1 that there exist vertices $v_1, v_2, \alpha \in V$ such that v_1 and v_2 are not adjacent, and $v_1 \rightarrow \alpha \leftarrow v_2$ in one graph, but in the other

$$v_1 \rightarrow \alpha \rightarrow v_2, \quad v_1 \leftarrow \alpha \rightarrow v_2 \quad \text{or} \quad v_1 \leftarrow \alpha \leftarrow v_2.$$

Suppose without loss of generality that $v_1 \to \alpha \leftarrow v_2$ in $\mathcal{D}_1$. Then in $\mathcal{D}_{1\cup 2}$ either $v_1 — \alpha$ or $\alpha — v_2$ (or both).

Hence, for any distribution in $M(\mathcal{D}_{1\cup 2})$, it must hold that for some set T with $\alpha \in T$, we have $v_1 \perp\!\!\!\perp v_2 \mid T$. However, it is easy to construct a distribution in $M(\mathcal{D}_1)$ in which $v_1 \not\!\perp\!\!\!\perp v_2 \mid T$ for any set T containing α. Suppose for example that all variables take states 0 and 1. Then let

$$\begin{aligned} P(X_v = x_v \,|\, x_{\mathrm{pa}(v)}) &= 1/2 \text{ for all } v \in V \setminus \{\alpha\} \\ P(X_\alpha = 0 \mid x_{v_1}, x_{v_2}) &= 2^{-1} + (-3)^{-(x_{v_1}+x_{v_2}+1)}. \end{aligned}$$

This completes the proof. □

One might also consider a population which is a mixture of two subpopulations described by two non-Markov equivalent DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$. In general such a population will not be in $M(\mathcal{D}_{1\cup 2})$. See Spirtes (1995) for further discussion.

5.3 Ordered blocking of variables

An elementary property of chain graphs is that the chain components partition the variables, and may be ordered so that all edges between variables within the same component are required to be undirected, while edges between variables in different components are directed in accordance with the ordering.

Applied contexts often suggest such an ordered blocking of variables. For example:

- variables may be divided into *Risk Factors*, *Diseases* and *Symptoms*;
- in a longitudinal study variables may be grouped according to time;
- in a cross-sectional study, causal knowledge may lead us to divide the variables into purely explanatory variables, intermediate variables and responses (Cox and Wermuth 1996).

Traditionally, such a *substantive* ordered blocking has been argued to justify modelling the variables via a chain graph with chain components compatible with the blocks, and with directed edges in accordance with the substantive ordering. (Wermuth and Lauritzen 1990; Whittaker 1990; Cox and Wermuth 1996). Below we show that in many contexts this procedure is incompatible with the goal of finding the most parsimonious independence model.

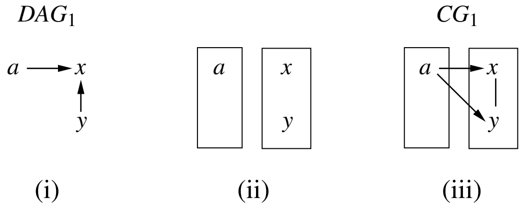


Figure 4: Restricting to chain graph models in keeping with a block ordering may lead to less parsimonious models: (i) DAG_1, the generating process; (ii) a block ordering $\{a\} \prec \{x, y\}$; (iii) CG_1 the minimal chain graph model for $\{a, x, y\}$, compatible with the ordering which contains the model given by DAG_1.

Suppose that it is known that a precedes x, but the relation between x and y is unknown, hence the blocking $\{a\} \prec \{x, y\}$ is proposed, as displayed in Figure 4 (ii) and that, in fact, the simple causal DAG_1 in Figure 4 (i) represents the true model.

The minimal chain graph on $\{a, x, y\}$ that is compatible with the blocking and contains the set of distributions over $\{a, x, y\}$ given by DAG_1 is saturated, as shown in in Figure 4 (iii). Thus a search for a chain graph model compatible with this blocking would not identify the simpler model given by DAG_1.

Consequently, leaving interpretation aside, restricting attention to chain graph models with a particular pre-specified blocking may preclude finding the most parsimonious model. It is also simple to see that had a, x and y been blocked together, the marginal independence would again be missed.

In the example just considered there were no unmeasured 'confounding' variables or selection variables. We now consider the case where such variables may be present, in the simple case of chain graphs with three vertices, but with one missing edge. Let $V = \{x_1, x_2, z\}$, with the missing edge occurring between x_1 and x_2. Up to symmetry of labelling x_1 and x_2, there are six different ways in which x_1, x_2 may be ordered relative to z, as indicated in the second column of Table 1: $v \sim w$ indicates that v and w are in the same component, while $v \prec w$ indicates that the component containing v precedes the component containing w in the ordering. Note that for (2) and (6) nothing is stated about the relation between the components containing x_1 and x_2, hence $x_1 \sim x_2$, $x_1 \prec x_2$ and $x_1 \succ x_2$ are all possible in these cases.

Table 1: Chain graphs with three vertices and two edges.

	Ordering	Edges in chain graph	Independence implied
(1)	$x_1 \sim z \sim x_2$	$x_1 — z — x_2$	
(2)	$x_1 \succ z \prec x_2$	$x_1 \leftarrow z \rightarrow x_2$	
(3)	$x_1 \sim z \prec x_2$	$x_1 — z \rightarrow x_2$	$x_1 \perp\!\!\!\perp x_2 \mid z$
(4)	$x_1 \prec z \sim x_2$	$x_1 \rightarrow z — x_2$	
(5)	$x_1 \prec z \prec x_2$	$x_1 \rightarrow z \rightarrow x_2$	
(6)	$x_1 \prec z \succ x_2$	$x_1 \rightarrow z \leftarrow x_2$	$x_1 \perp\!\!\!\perp x_2$

The edges between x_1 and z, and x_2 and z, are then determined by the ordering, and take the form shown. It then follows from the global Markov property for chain graphs (Lauritzen 1996, p. 55) that in cases (1) to (5) $x_1 \perp\!\!\!\perp x_2 \mid z$, while in case (6) $x_1 \perp\!\!\!\perp x_2$.

We will show by example that for each of the orderings specified in Table 1 there exist DAGs containing x_1, x_2 and z which obey the specified ordering, and yet violate the conditional independence relations specified by a chain graph under this ordering.

For cases (1) to (5) of Table 1 consider DAG_3 shown in Figure 5 (i), in which h_1 and h_2 are *hidden* (i.e. unobserved) variables. It is easy to see that this generating process is compatible with *any* of the orderings given in column 2 of Table 1. However, under this model $x_1 \perp\!\!\!\perp x_2$, contradicting the independence implied by a chain graph under the block ordering.

For case (6) consider DAG_4 shown in Figure 5. Whereas in DAG_3, we marginalized h_1 and h_2, here we consider $P(x_1, x_2, z \mid s)$. A simple interpretation of s, is that it represents a *selection* variable, which takes the same value for all units in the sub-population being modelled. See Cox and Wermuth (1996), p. 44; Cooper (1995); Spirtes *et al.* (1995), Spirtes and Richardson (1997) and Lauritzen (1999) for further discussion. In the conditional distribution $P(x_1, x_2, z \mid s)$ it holds that $x_1 \perp\!\!\!\perp x_2 \mid z$, rather than marginal independence which is implied by the chain graph under this ordering.

In fact it may be shown that any DAG in which $x_1 \perp\!\!\!\perp x_2 \mid z$ with the ordering given by (6) will contain variables that are conditioned on; marginalization alone is not sufficient. This may explain why it is often

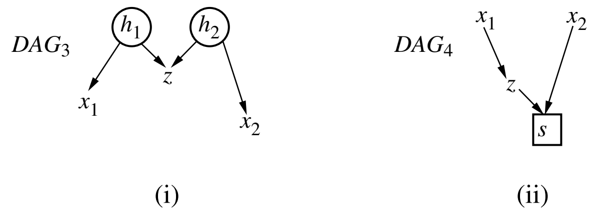


Figure 5: Examples showing that, when latent or selection variables may be present, ordering of variables does not imply the conditional independence relations given by a chain-graph. In DAG_3 h_1 and h_2 are hidden variables. In DAG_4 s is a selection variable that has been conditioned on.

inferred that conditional independence given z is incompatible with $x_1 \prec z \succ x_2$. See e.g. Mohamed *et al.* (1998), p. 353.

There are, of course, generating processes which simultaneously satisfy the ordering and conditional independence structures given in Table 1. Still, we conclude that without background knowledge that rules out hidden, confounding variables, or selection effects, or rules in the presence of certain edges, information on the ordering of variables cannot be used reliably to infer conditional independence structure. Hence knowledge of an ordered blocking of variables alone is not sufficient to justify postulating a chain graph compatible with those blocks; additional detailed substantive arguments, ruling out (or hypothesizing) the absence of confounders are always required.

We conclude this section by making several further points:

- The chain graphs in the examples given contained at most three vertices. If we view these graphs as induced subgraphs of a larger chain graph, then the whole discussion carries over if instead of $x_1 \perp\!\!\!\perp x_2 \mid z$, and $x_1 \perp\!\!\!\perp x_2$ we consider $x_1 \perp\!\!\!\perp x_2 \mid W$, with $z \in W$ and $z \notin W$ respectively.

- The problems we have highlighted that arise due to the presence of hidden variables, would still be present even if all chain components were singletons, i.e. if we considered DAGs under a fixed ordering.

- There are independence structures arising from DAGs with hidden variables that cannot be represented by any chain graph model. Figure 2 (i) is an example. Richardson and Spirtes (2000) provide a graphical representation of these structures. However, in the simple cases involving three vertices there is always a chain graph representing the independence structure. This raises the question why not, in such circumstances, just ignore the blocking and represent the independence structure directly?

- Often it appears that resistance to consideration of models that violate blocking follows from a naive causal interpretation of the resulting graph. Thus for instance, if DAG_3 in Figure 5 (i) is the generating process, then the independence structure can be represented by the chain graph $x_1 \to z \leftarrow x_2$. However, if the variables are ordered, e.g. by time, as $z \prec \{x_1, x_2\}$ then such a model appears to represent the absurdity of the future causing the past. However, if regarded strictly as representing an independence hypothesis then such a model presents no difficulties: in fact, it would lead us to the (correct) conclusion that unmeasured confounding variables are present. Sticking to the blocking would conceal the marginal independence of x_1 and x_2.

- In some cases, more principled objections to consideration of a less restricted class of chain graphs may be adduced: there may be computational issues involved in searching a larger model class, or there may be an intuition that it is unwise to consider too rich a model class if data is insufficient. However, it would have to be argued that in these respects, a particular class of chain graphs was superior to simple undirected graphs.

6 Feedback Models for Chain Graphs

As demonstrated by the previous discussion, chain graph models represent qualitatively different hypotheses from those represented by DAG models, including DAG models under marginalization and conditioning. This suggests that any general data generating process for chain graph models must involve infinite processes converging to some type of equilibrium.

In this section we present a number of alternative equilibrium data generating processes with feedback that all lead to chain graph models.

We first consider the special case of an undirected graph $\mathcal{G}$ and an associated distribution P with positive density f which factorizes according to the graph, i.e. it has the form

$$f(x) = \prod_{c \in \mathcal{C}} \phi_c(x), \tag{6}$$

where ϕ_c depends on x through x_c only and $\mathcal{C}$ denotes the set of cliques (maximal complete subsets) of $\mathcal{G}$. The original idea behind such graphical models originates in statistical physics (Gibbs 1902), where x denotes possible states of a physical system and $f(x)$ is proportional to $\exp\{-E(x)\}$ with $E(x)$ denoting the total energy of the system in state x. The energy is then assumed to be additively built up by *potentials* ψ_c as,

$$E(x) = \sum_c \psi_c(x_c) = -\sum_c \log \phi_c(x).$$

There are several alternative dynamic systems that all have the distribution P as their equilibrium distribution. This has been extensively exploited in the literature on Markov chain Monte Carlo methods for simulating from P (Metropolis *et al.* 1953; Hastings 1970; Geman and Geman 1984; Gilks *et al.* 1996). We describe a few of these dynamic regimes below.

Note that the dynamic regimes apply to any distribution with positive density, the only use of the factorization (6) is to simplify computations during updating.

6.1 Data generating processes for undirected graphs

6.1.1 The systematic Gibbs sampler

The dynamic regime which is simplest to explain is based on the *systematic Gibbs sampler* which evolves in discrete time and proceeds by choosing an arbitrary value $x^0 \in \mathcal{X}$ and an arbitrary ordering of the vertices in V so that $V = \{1, \ldots, p\}$. The vertices are then visited in the given order, each X_v being updated according to its conditional distribution given the values of X at the remaining vertices. The factorization (6) implies that the density of this conditional distribution simplifies as

$$f(x_i \,|\, x_{-i}) = f(x_i \,|\, x_{\mathrm{bd}(i)}) \propto \prod_{c: i \in c} \phi_c(x),$$

where x_{-i} is a short notation for $x_{V \setminus \{i\}}$. The corresponding generating process can in an idealized form be written as the following 'computer program':

$x \leftarrow x^0$;

$i \leftarrow 0$;

`Repeat until equilibrium:`

$i \leftarrow i + 1 \bmod p$;

$x_i \leftarrow y_i$ `with probability` $f(y_i \mid x_{-i})$;

`Return` x.

The (random) output X_τ of this program will have distribution P as desired.

The expressions '`until equilibrium`' and '`Return` x' have to be understood in the way that the random assignments are repeated a very large number of times, so that a 'stochastic' equilibrium prevails and then the program returns a 'shapshot' in time of the configurations of the variables.

The system involves feedback in the sense that the value of X_i for any $i \in V$ has been dynamically affected by all of the variables $X_{\mathrm{bd}(i)}$.

6.1.2 The random Gibbs sampler

The *random Gibbs sampler* proceeds in a similar way, only the variable to be updated is chosen at random. Thus here we need not order the variables and can write the corresponding program as

$x \leftarrow x^0$;

`Repeat until equilibrium:`

$v \leftarrow$ `rand`(V);

$x_v \leftarrow y_v$ `with probability` $f(y_v \mid x_{-v})$;

`Return` x.

where rand(V) chooses a random element from the set V.

6.1.3 Time reversible Markov dynamics

This dynamic regime applies to the case of a discrete state space and is in many ways physically more plausible than the discrete time schemes described above.

Here the system is assumed to develop as a Markov process in continuous time with intensities of the form

$$P\{X(t+dt) = y \mid X(t) = x\} = \begin{cases} q_v(y_v, x)dt + o(dt) & \text{if } y = (y_v, x_{-v}), \text{ and } y_v \neq x_v \\ 1 - q(x)dt + o(dt) & \text{if } y = x \\ o(dt) & \text{otherwise} \end{cases}$$

with q_v having $q_v(x_v, x) = 0$ and $q(x) < 1$, where $q(x) = \sum_v \sum_{y_v \neq x_v} q_v(y_v, x)$. If q_v is suitably chosen, these equations describe a time-reversible Markov process with P as equilibrium distribution (Spitzer 1971; Preston 1973; Besag 1974a).

In this dynamic model, the system is at rest for an exponentially distributed length of time and then a randomly chosen site is updated as before. The distribution of the waiting time depends in general on the current configuration of the system and this is also true of the conditional distribution of the site to be updated.

6.1.4 Langevin diffusions

In the case of a continuous state space with smooth densities, there is an alternative and very simple diffusion process known as the *Langevin* diffusion given as

$$X(t + dt) = X(t) + \frac{1}{2} \operatorname{grad} \log f\{X(t)\}\, dt + dW(t) \tag{7}$$

where W is standard $|V|$-dimensional Brownian motion. Under suitable smoothness conditions on f (Roberts and Tweedie 1996), this dynamic scheme also has P as an equilibrium distribution. This has, for example, been exploited by Grenander and Miller (1994). Also here, the gradient simplifies due to the factorization (6), we omit the details.

6.1.5 The Gaussian case

Next we consider the special case when the joint distribution is assumed to be multivariate Gaussian with mean zero and a regular covariance matrix Σ with inverse $K = \Sigma^{-1}$. If the distribution satisfies the Markov property of an undirected graph, we have

$$k_{uv} = 0 \text{ whenever } u \not\sim v. \tag{8}$$

Gibbs dynamics If the vertices of the graph are numbered as $V = \{1, \ldots, p\}$, the Gibbs dynamics is also known as *conditional autoregression* (CAR) (Ripley 1981) or an *auto-normal prescription* (Besag 1975). Here at time t each variable is updated linearly as

$$x_v \leftarrow \sum_{u:u \neq v} a_{vu} x_u + \epsilon_v$$

where ϵ_v is distributed as $\mathcal{N}(0, 1/k_{vv})$ and $a_{vu} = -k_{vu}/k_{vv}$. If the distribution satisfies the Markov property of an undirected graph, (8) implies that the sum above only extends over the neighbours of v. We will write this dynamic scheme as

$$X(t+1) \stackrel{G}{\longleftarrow} A * X(t) + \epsilon(t+1) \tag{9}$$

where A is the matrix of coefficients. The special assignment symbol and asterisk indicates that this is not a standard matrix equation but updating is made sequentially by row.

Clearly, although any matrix A would make sense in the updating equation (9), it would not necessarily correspond to Gibbs updating for a multivariate Gaussian distribution with some covariance matrix Σ. For this to be the case, A must at least have zero diagonal elements and also satisfy an equation of *balance*

$$a_{uv}\sigma_{vv} = a_{vu}\sigma_{uu}, \tag{10}$$

where σ_{vv} is the variance of the innovation $\epsilon_v(t)$. If the variables are scaled to have unity innovation variances, the necessary and sufficient condition for the CAR system to be a Gibbs updating scheme corresponding to a multivariate Gaussian distribution is that A have zero diagonal elements and that $I - A$ be symmetric and positive definite (Besag 1975; Ripley 1981). The covariance matrix of the equilibrium distribution is then given as $\Sigma = (I - A)^{-1}$.

The equation (9) can be reexpressed in terms of a *simultaneous autoregression* (SAR) as

$$X(t+1) \leftarrow BX(t) + U(t+1),$$

where B and U are determined appropriately from A and ϵ.

Langevin dynamics In the Gaussian case, the Langevin diffusion corresponds to the stochastic differential equation

$$X(t+dt) = X(t) - \frac{1}{2}KX(t)\,dt + dW(t). \tag{11}$$

Besag (1974a) studied Markov systems as equilibrium distributions for more general diffusions of the type

$$X(t+dt) = X(t) + CX(t)\,dt + dZ(t), \tag{12}$$

where $Z(t)$ is Brownian motion with covariance matrix $\mathbf{V}\{dZ(t)\} = \Lambda$, see also Cox and Wermuth (2000). The equilibrium distribution exists if and only if C is a stability matrix, i.e. the real parts of the eigenvalues of C are negative. In this case, the equilibrium distribution is determined as the Gaussian distribution with mean zero and covariance matrix equal to the unique solution of the matrix equation

$$\Lambda + C\Sigma + \Sigma C^\top = 0. \tag{13}$$

Clearly there are many more choices for C and Λ leading to $\Sigma = K^{-1}$ than the choice $C = -K/2$ used in the Langevin diffusion (11).

6.2 Intervention in undirected graphs

Each of the dynamic schemes described above correspond in a natural way to intervention models. For the systematic and random Gibbs sampler as well as the time-reversible Markov dynamics, the intervention $X_A \leftarrow x_A^*$ corresponds to replacement of the corresponding lines in the program, just as in the DAG case. Clearly, when intervention is carried out in this way, it has the same effect as ordinary conditioning, i.e. for $B = V \setminus A$, we have

$$P(X_B = x_B \,|\, X_A \leftarrow x_A^*) = P(X_B = x_B \,|\, X_A = x_A^*). \tag{14}$$

For the Langevin dynamics, the natural description of the effect of an intervention $X_A \leftarrow x_A^*$ would be to replace the original diffusion equation (7) with

$$X_B(t + dt) = X_B(t) + \frac{1}{2} \operatorname{grad} \log f\{X_B(t), x_A^*\}\, dt + dW_B(t). \tag{15}$$

Since the density obtained by conventional conditioning is given as

$$f(x_B \,|\, x_A^*) \propto f(x_B, x_A^*),$$

the diffusion (15) has equilibrium equal to this conditional distribution, so that (14) also holds in this case.

Note that if we consider a more general dynamic regime such as the diffusion (12) this may no longer be true. The effect of an intervention under this diffusion leads naturally to

$$X_B(t + dt) = X_B(t) + C_{BB} X_B(t)\, dt + C_{BA} x_A^*\, dt + dZ_B(t), \tag{16}$$

where the matrix C has been partitioned into appropriate blocks. Indeed, in the Gaussian case it holds that if the intervention diffusion always has equilibrium distribution equal to the conditional distribution, it must be the Langevin diffusion, which is seen as follows:

The equilibrium distribution of the intervention diffusion (16) has expectation equal to

$$\mathbf{E}(X_B \,||\, x_A^*) = -C_{BB}^{-1} C_{BA} x_A^*$$

and its covariance matrix is the unique symmetric solution Ω_{BB} to the equation

$$I + C_{BB}\Omega_{BB} + \Omega_{BB} C_{BB}^\top = 0,$$

where we assume that the variables are scaled to have unity innovation variances.

If this distribution is equal to the conditional distribution, we have

$$C_{BB}^{-1} C_{BA} = K_{BB}^{-1} K_{BA} \tag{17}$$

and

$$I + C_{BB} K_{BB}^{-1} + K_{BB}^{-1} C_{BB}^\top = 0. \tag{18}$$

From the special case where $B = \{v\}$ is a singleton, we obtain from (18)

$$c_{vv} = -k_{vv}/2$$

and inserting this into (17) yields for all $u \neq v$

$$c_{vu} = c_{vv} k_{vv}^{-1} k_{vu} = -k_{vu}/2$$

and thus $C = -K/2$ as required.

6.3 Data generating processes for chain graphs

We recall from Section 3.1 that in a chain graph situation, we have a distribution P with a density which factorizes in two stages (Lauritzen 1996). If $\mathcal{T}$ denotes the set of chain components of $\mathcal{G}$, we have

$$f(x) = \prod_{\tau \in \mathcal{T}} f(x_\tau \,|\, x_{\mathrm{pa}(\tau)}),$$

where each factor further factorizes according to the graph $\mathcal{G}^*(\tau)$.

Similarly, the data generating processes for chain graph models have two loops. The outer loop corresponds to the DAG of chain components, where each chain component is updated in a scheme satisfying the restriction that variables in parent components have been assigned their values when the update is to be made:

$$X_\tau \leftarrow G_\tau(X_{\mathrm{pa}(v)}), \tau \in \mathcal{T}.$$

The inner loop, represented by G_τ, updates the variables in the chain component τ. For those components that are not singletons, G_τ represents one of the generating processes for undirected graphs applied to a chain component τ for a fixed value of the variables at its parents $x_{\mathrm{pa}(\tau)}$. It then becomes a function of these, so that the program G_τ takes $x_{\mathrm{pa}(\tau)}$ as input and gives x_τ as output. In its random form, the program becomes

```
Function G_τ;
    input x_pa(τ);
    x_τ ← x_τ^0;
    Repeat until equilibrium:
        v ← rand(τ);
        x_v ← y_v  with probability  f(y_v | x_τ\{v}, x_pa(τ));
Return x_τ
```

and similarly in its systematic form. Note that only variables in the specific chain component τ are updated during this inner loop. Thus variables on an 'equal footing' are updated in the same inner loop if they are also in the same chain component, whereas such variables are updated independently and possibly in parallel if they are in the same 'box' but different chain components.

The above procedure can be written in a way that makes its functional character more explicit, thereby making the analogy to traditional structural equation systems clearer. We let $\epsilon^\tau = (\epsilon^1, \epsilon^2, \ldots)$ denote a (potentially infinite) sequence of independent and identically uniformly distributed variables which are used as input to the function g_τ, jointly with $x_{\mathrm{pa}(\tau)}$. Again using the random variant of the Gibbs sampler, this yields

```
Function G_τ;
    input (x_pa(τ), ε^τ);
    x_τ ← x_τ^0;
    n ← 0;
    Repeat until equilibrium:
```

$v \leftarrow$ `rand`(τ);
$n \leftarrow n + 1$;
$x_v \leftarrow h_v^\tau(x_{\tau\setminus\{v\}}, x_{\mathrm{pa}(\tau)}, \epsilon^n)$;

`Return` x_τ.

Here h_v^τ is chosen so that if U is uniformly distributed on the unit interval, then $h_v^\tau(x_{\tau\setminus\{v\}}, x_{\mathrm{pa}(\tau)}, U)$ has density $f(y_v \mid x_{\tau\setminus\{v\}}, x_{\mathrm{pa}(\tau)})$, i.e. h_v^τ is a direct Monte Carlo simulator for this conditional distribution.

If the chain component τ is a singleton, equilibrium is achieved immediately, and we simply get that

$$g_\tau(x_{\mathrm{pa}(\tau)}, \epsilon^\tau) = h^\tau(x_{\mathrm{pa}(\tau)}, \epsilon^1).$$

If we order the chain components as $\tau_1, \ldots, \tau_p$ and the variables in each chain component $\tau_i = \{n_i + 1, \ldots, n_i + t_i\}$ and use the systematic variant of the Gibbs sampler, a full structural assignment system associated with a general chain graph has the form

$x \leftarrow x_0$;

`For` $i = 1, \ldots, p$

$j \leftarrow 0$;

`Repeat until equilibrium:`

$j \leftarrow j + 1 \bmod t_i$
$x_{n_i+j} \leftarrow h_i^j(x_{\tau_i\setminus\{j\}}, x_{\mathrm{pa}(\tau)},$ `runif`$)$;

`Return` x;

where again h_i^j is suitably chosen. As in the directed acyclic case, we have that

Proposition 5 *If P is a distribution with strictly positive density which satisfies the Markov property on the chain graph $\mathcal{G}$ and X is defined through a structural assigment system as above, then X has distribution P.*

Proof The fact that the structural assignment system leads to $(X_\tau, \tau \in \mathcal{T})$ satisfying the Markov property of the DAG formed by the chain components of $\mathcal{G}$ is seen exactly as in the directed acyclic case, see for example Lauritzen (2001), Theorem 2.20.

Clearly, for each fixed $x_{\mathrm{pa}(\tau)}$, the conditional distribution of the random function $G_\tau(x_{\mathrm{pa}(\tau)})$ has density $f(x_\tau \mid x_{\mathrm{pa}(\tau)})$ as the Gibbs sampler was designed to sample the variables in τ from this conditional distribution. Thus the joint density of X must be given by (2) as desired. □

We have thus constructed a number of dynamic regimes which all lead to models with conditional independence structure determined by a chain graph.

Since equilibrium will never be attained, each of the generating processes is to be considered an approximation to a situation in which the real updating within each chain component is developing so fast that the equilibrium can be considered instantaneous, relative to the time elapsed between generation of different chain components. Each chain component outputs a random snapshot of its state, which in turn is used as input for the next chain component equilibrium process.

The plausibility of such generating processes in any given context clearly depends on that context. Generally, systematic updating seems somewhat unnatural as there cannot be a natural ordering of variables considered 'on an equal footing' and the more complex schemes of random updating, continuous time Markov processes, or diffusions, have generally more intuitive appeal.

6.4 Intervention in chain graphs

If the intervention $X_A \leftarrow x_A$ is made by replacement in each chain component as described in Section 6.2, it follows as in the directed case that this leads to the the formula

$$p(x \,||\, x_A) = \prod_{\tau \in \mathcal{T}} p(x_{\tau \setminus A} \mid x_{\mathrm{pa}(\tau)}, x_{\tau \cap A}). \tag{19}$$

This formula specializes to the intervention formula (5) in the fully directed case and Bayes' formula in the undirected case. It also corresponds to the analogy with decision networks based on chain graphs as discussed in Cowell *et al.* (1999), where interventions are then described by decision nodes.

An alternative argument for (19) may be based on the assumption that the potentials $\psi_A = \log \phi_A$ from (3) are stable under intervention, as they represent physical laws beyond control of the intervening. This directly generalizes the idea used for causal DAGs, where conditional distributions of children given parents were considered stable under intervention.

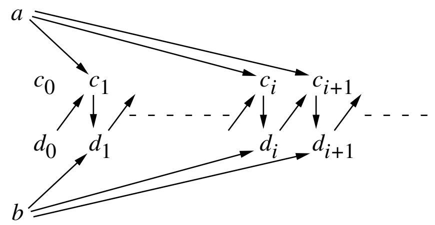


Figure 6: Infinite DAG corresponding to a structural assignment system for CG_3 of Figure 1 where c is updated before d in each inner loop.

6.5 Equilibrium dynamics and infinite DAGs

It is illuminating to think of the equilibrium dynamics described in terms of infinite DAGs. If we, for example, consider the simple chain graph CG_3 in Figure 1 (ii), the generating process corresponding to this graph using the systematic Gibbs sampler dynamics would first independently choose values x_a and x_b for the variables labelled a and b, then use these as input for an equilibrium process updating of c and d as indicated in Figure 6. Using d-separation on the DAG in Figure 6 yields

$$d_i \perp\!\!\!\perp a \mid \{c_i, b\} \text{ and } c_i \perp\!\!\!\perp b \mid \{d_{i-1}, a\}$$

whereas in general

$$c_i \not\perp\!\!\!\perp b \mid \{d_i, a\}$$

since b and c_i are common parents of d_i in the update scheme described. Thus taking a snapshot as

$$(X_c, X_d) \leftarrow (X_{c_i}, X_{d_i})$$

will not reproduce the desired conditional independence $c \perp\!\!\!\perp b \mid \{d, a\}$.

However, when the conditional distributions in the infinite DAG are consistent in the sense that for fixed values (x_a, x_b) there is a joint distribution of (X_c, X_d) from which the conditional update distributions are derived (as hold under Gibbs dynamics), then (X_{c_i}, X_{d_i}) and $(X_{c_i}, X_{d_{i-1}})$ have the same equilibrium distribution. It therefore holds *in equilibrium* and thus approximately for large i that $c_i \perp\!\!\!\perp b \mid \{d_i, a\}$, provided such update distributions are used.

7 Linear Structural Equation Models

7.1 Basic terminology

In a linear structural equation model (SEM), variables are conventionally divided into two disjoint sets: substantive variables and error variables. Associated with each substantive variable $X_v, v \in V$ there is a unique error term ϵ_v. A linear SEM contains a set of linear equations, one for each substantive variable, expressing X_v as a linear function of the other substantive variables, together with ϵ_v. In vector notation:

$$X = \Gamma X + \boldsymbol{\epsilon},$$

where $\gamma_{vv} = 0$. In any given structural model some off-diagonal entries in Γ may also be fixed at zero, depending on the form of the structural equations. If under some rearrangement of the rows, Γ can be placed in lower triangular form, the system of equations is said to be *recursive*, otherwise it is said to be *non-recursive*.

If we define a directed graph with vertex set V by having a directed edge from u to v if and only if γ_{vu} is not fixed at zero, a SEM is recursive precisely when this graph is a DAG. In a non-recursive system, there might be edges between vertices in both directions if γ_{uv} and γ_{vu} are both allowed to be non-zero.

In addition a SEM model specifies a multivariate normal distribution over the error terms: $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Delta)$. In any particular model, some off-diagonal (δ_{ij}) entries in Δ may be specified to be non-zero. If Δ is not diagonal then the model is said to have *correlated errors*.

If $(I - \Gamma)$ is non-singular, the SEM determines a joint distribution over the substantive variables, which can be derived from the *reduced form* equations:

$$X = (I - \Gamma)^{-1}\boldsymbol{\epsilon},$$

yielding

$$X \sim \mathcal{N}(0, \Sigma) \text{ with } K = \Sigma^{-1} = (I - \Gamma)^{\top}\Delta^{-1}(I - \Gamma).$$

This should be contrasted with the CAR interpretation (9) which in the case of $\Delta = I$, $A = \Gamma$, and $I - \Gamma$ positive definite would lead to $K = (I - \Gamma)$.

The following example of a non-recursive SEM with uncorrelated errors which can naturally be associated with the directed graph of Figure 2 (iii) with a relabelling of the vertices:

$$\begin{aligned} x_1 &= \epsilon_1 \\ x_2 &= \epsilon_2 \\ x_3 &= \gamma_{31}x_1 + \gamma_{34}x_4 + \epsilon_3 \\ x_4 &= \gamma_{42}x_2 + \gamma_{43}x_3 + \epsilon_4 \end{aligned} \qquad \Delta = \begin{pmatrix} \delta_{11} & 0 & 0 & 0 \\ 0 & \delta_{22} & 0 & 0 \\ 0 & 0 & \delta_{33} & 0 \\ 0 & 0 & 0 & \delta_{44} \end{pmatrix}.$$

As mentioned in Section 4.2, we would generally prefer to use the term 'structural assignment model' but have chosen to stick with the more traditional terminology throughout this section.

Fisher (1970) presents a dynamic process, whose time average gives the distribution described by a linear non-recursive structural equation model. Here the system is occasionally subjected to random exogeneous disturbances of the exact equilibrium. The eigenvalues of Γ are required to be less than 1 for convergence of the time averages, see Richardson (1996) for a more detailed description of this equilibrium process.

Using the intervention interpretation of structural equations given by Strotz and Wold (1960) leads for the non-recursive case to an intervention distribution which is quite different from those earlier described. Indeed, if in the example given we intervene as $X_4 \leftarrow x_4^*$ we obtain the recursive SEM

$$\begin{aligned} x_1 &= \epsilon_1 \\ x_2 &= \epsilon_2 \\ x_3 &= \gamma_{31}x_1 + \gamma_{34}x_4^* + \epsilon_3 \end{aligned} \qquad \Delta = \begin{pmatrix} \delta_{11} & 0 & 0 \\ 0 & \delta_{22} & 0 \\ 0 & 0 & \delta_{33} \end{pmatrix}. \tag{20}$$

7.2 Chain graph models for structural equations

The chain graph models and corresponding dynamics described can in some cases give an alternative interpretation of a structural equation system with coefficient matrix Γ. As opposed to the interpretation above which can be seen as a deterministic equilibrium with random boundary conditions, we then get an interpretation in terms of a stochastic equilibrium.

To make such an interpretation we associate an undirected edge with every pair (u, v) for which γ_{uv} and γ_{vu} are allowed to have non-zero values, instead of two directed edges as used above. Thus, the SEM described in the above example would then correspond to the graph CG_3 in Figure 1 (ii).

The graph of a SEM under this interpretation may not in general be a chain graph and unless this is the case, the model will not have a chain graph interpretation. But if this graph happens to be a chain graph, the dynamic schemes discussed in Section 6 could be used to give an alternative interpretation of a structural equation model with feedback. Thus in each chain component of the graph, the structural equations could be implemented as conditional autoregressions (9). As mentioned in Section 6.1.5, such a specification does not always correspond to a well-defined joint distribution. The system should satisfy

$$\gamma_{uv}\delta_{vv} = \gamma_{vu}\delta_{uu} \text{ whenever both are non-zero} \tag{21}$$

and in addition — if we again assume the variables have been scaled to have unity error variances — the submatrix of $I - \Gamma$ induced by the corresponding chain component would have to be positive definite. In the example considered above, these conditions would amount to having

$$\gamma_{34}\delta_{44} = \gamma_{43}\delta_{33} \text{ and } \gamma_{34}\gamma_{43} < 1.$$

The first condition ensures balance whereas the second condition ensures stability of the dynamic system.

Thus, non-recursive structural equation models would only admit a chain graph representation under quite special circumstances and the 'equal footing' of variables in the same chain component under this interpretation demands complete 'symmetry of forces' as represented by the relation (21).

If the conditions above are fulfilled, the distribution after intervention as $X_4 \leftarrow x_4^*$ becomes the same as in (20), but now it is obtained from the joint distribution by the intervention formula (19). The joint distribution is different under the chain graph interpretation of the SEM, for which (19) would not lead to the distribution (20).

Ord (1976) also suggested use of the CAR interpretation for simultaneous equation models in economics whereas Wermuth (1992) suggested a quite different chain graph representation of a structural equation system with feedback which also demands special restrictions on the parameters, see also Lauritzen (1996) pp. 154–155.

8 Discussion

The results presented in this paper have consequences in several contexts:

8.1 Causal DAGs versus causal chain graphs

As described in Section 4, there is a large body of work which takes as its starting point the assumption that the population of interest is described by a causal DAG possibly with some variables unobserved. The considerations in Section 6 indicate that in some circumstances this assumption may be unduly restrictive: if feedback is present then the model for the population of interest could sometimes be adequately described by a causal chain graph. See also Bentzel and Hansen (1954) for a similar discussion in the context of recursive vs. non-recursive SEMs.

8.2 Undirected edges and causal under-determination

As mentioned in Section 5, one original motivation for introducing graphs with both undirected and directed edges was to allow direct associations that were not assumed to be causal. In particular an analysis which leads to a chain graph, rather than a DAG, might at first sight appear to be more 'causally prudent'. However, as we have shown, the situation is more complicated:

- If the chain graph is not Markov equivalent to a recursive 'causal' graph, then the graph contains an undirected edge which essentially is *only* interpretable via feedback.
- Chain graphs do not in general represent the independence structures that arise from DAGs with hidden variables. For this purpose, ancestral graphs are required.
- A chain graph may be used to represent the union of a set of DAGs with common adjacencies only if the DAGs are all Markov equivalent.

Thus only certain undirected edges may be interpreted as (prudently) representing a collection of causal hypotheses. Further, there are alternate causal hypotheses involving hidden variables that are excluded by restricting attention to chain graphs.

8.3 Data analyses using chain graph models and blocking

As shown in section 5.3, restricting attention to the class of chain graphs compatible with a pre-specified ordering will often be incompatible with finding the most parsimonious model. This seems undesirable:

- If the primary goal of the analysis is prediction (of the joint distribution) then parsimonious models are often preferable.
- If explanation is the goal then a less parsimonious model — which will include 'extra' edges — may often be misleading.

However, if the goal is to gain insight into possible causal data generating processes then the most parsimonious model may fail to represent all causal relations if there is *parametric cancellation* — also known as a 'violation of faithfulness' in the terminology of Spirtes *et al.* (1993) — since in this case not all the independence relations holding in the population will be due to causal structure. In many circumstances it may be reasonable to assume that such cancellations do not occur (Spirtes *et al.* 1993; Meek 1995; Pearl 2000), but without such an assumption, the most parsimonious model will not reflect the process that generated the data. On the other hand, if one has good reason to believe parametric cancellation is present, then this might argue against attempting to model the independence structure in order to understand the generating process.

If background knowledge is available it would seem desirable to exploit this when performing model determination. However, as shown in Section 5.3, when hidden variables may be present, knowledge about ordering may not yield any information which is relevant for restricting the class of possible independence models. An alternative approach would be to use background knowledge *after* model determination has been completed in order to narrow down a set of candidate models.

8.4 Chain graphs under the alternative Markov property

An alternative Markov property for chain graphs has been proposed by Andersson *et al.* (1996). Hence, in general, there are different statistical models that may be associated with the same chain graph. For example, within this alternative Markov property the graph CG_3 in Figure 1 (ii) implies the independence relations

$$a \perp\!\!\!\perp b \qquad a \perp\!\!\!\perp \{b, d\} \qquad b \perp\!\!\!\perp \{a, c\}$$

and hence this model is Markov equivalent to the generating process corresponding to DAG_4 in Figure 2 (i). However, there are other chain graphs for which the alternative property results in an independence model that again cannot be obtained from any finite DAG by marginalizing or conditioning (Richardson 1998).

In this paper we have shown that chain graphs under the original Markov property describe certain types of feedback system. This naturally raises the question as to which generating processes correspond to chain graphs under this alternative Markov property. Cox and Wermuth (1993) discuss other possible interpretations of chain graphs, for which the same question may arise.

8.5 Summary

A remark in Spiegelhalter *et al.* (1993), foreshadows many of our conclusions: in a response to comments made by Glymour and Spirtes they state that "chain graph models represent... equilibrium systems" (p. 278). In this paper we have constructed dynamic processes with equilibria corresponding to chain graphs, and we have also shown that this remark may be strengthened to say that, in general, chain graph models *only* represent such systems well and then under quite subtle dynamic regimes. In addition, we have extended the intervention theory for DAGs to these dynamic systems.

References

Andersson, S. A., Madigan, D., and Perlman, M. D. (1996). An alternative Markov property for chain graphs. In *Uncertainty in Artificial Intelligence: Proceedings of the* 12^{th} *Conference*, (ed. F. V. Jensen and E. Horvitz), pp. 40–8. Morgan Kaufmann, San Francisco.

Bentzel, R. and Hansen, B. (1954). On recursiveness and interdependency in economic models. *Review of Economic Studies*, **22**, 153–68.

Besag, J. (1974a). On spatial-temporal models and Markov fields. In *Transactions of the* 7^{th} *Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 47–55. Academia, Prague.

Besag, J. (1974b). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 302–9.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**, 179–95.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* John Wiley and Sons, New York.

Cooper, G. F. (1995). Causal discovery from data in the presence of selection bias. In *Preliminary papers of the 5th International Workshop on AI and Statistics, January 4-7, Fort Lauderdale, Florida*, (ed. D. Fisher), pp. 140–50.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems.* Springer-Verlag, New York.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statistical Science*, **8**, 204–218; 247-277.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation.* Chapman and Hall, London.

Cox, D. R. and Wermuth, N. (2000). On the generation of the chordless four-cycle. *Biometrika*, **87**, 204–12.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, **8**, 522–39.

Edwards, D. and Kreiner, S. (1983). The analysis of contingency tables by graphical models. *Biometrika*, **70**, 553–62.

Fisher, F. M. (1970). A correspondence principle for simultaneous equation models. *Econometrica*, **38**, (1), 73–92.

Frydenberg, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics*, **17**, 333–53.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, **6**, 721–41.

Gibbs, W. (1902). *Elementary Principles of Statistical Mechanics.* Yale University Press, NewHaven, Connecticut.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo Methods in Practice.* Chapman and Hall, New York.

Goldberger, A. S. (1972). Structural equation models in the social sciences. *Econometrica*, **40**, 979–2001.

Grenander, U. and Miller, M. I. (1994). Representations of knowledge in complex systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **56**, 549–603.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, **11**, 1–12.

Hammersley, J. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks.* University College London Press, London.

Kiiveri, H. and Speed, T. P. (1982). Structural analysis of multivariate data: A review. In *Sociological Methodology*, (ed. S. Leinhardt). Jossey-Bass, San Francisco.

Kiiveri, H., Speed, T. P., and Carlin, J. B. (1984). Recursive causal models. *Journal of the Australian Mathematical Society, Series A*, **36**, 30–52.

Koster, J. T. A. (1996). Markov properties of non-recursive causal models. *Annals of Statistics*, **24**, 2148–77.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford, United Kingdom.

Lauritzen, S. L. (1999). Generating mixed hierarchical interaction models by selection. Technical Report R-99-2021, Dept. of Mathematical Sciences, University of Aalborg, Aalborg, Denmark.

Lauritzen, S. L. (2001). Causal inference from graphical models. In *Complex Stochastic Systems*, (ed. O. E. Barndorff Nielsen, D. R. Cox, and C. Klüppelberg), pp. 63–107. Chapman and Hall/CRC Press, London/Boca Raton.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **50**, 157–224.

Lauritzen, S. L. and Wermuth, N. (1984). Mixed interaction models. Technical Report R 84-8, Institute for Electronic Systems, Aalborg University.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31–57.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, (ed. P. Besnard and S. Hanks), pp. 403–10. Morgan Kaufmann Publishers, San Francisco, CA.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–92.

Mohamed, W. N., Diamond, I., and Smith, P. W. F. (1998). The determinants of infant mortality in Malaysia: a graphical chain modelling approach. *Journal of the Royal Statistical Society, Series A*, **161**, 349–66.

Ord, K. (1976). An alternative approach to modelling linear systems. Unpublished manuscript.

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems.* Morgan Kaufmann Publishers, San Mateo, CA.

Pearl, J. (1993). Graphical models, causality and intervention. *Statistical Science*, **8**, 266–9. Comment to Spiegelhalter *et al.* (1993).

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669–710.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research*, **27**, 226–84.

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

Preston, C. J. (1973). Generalised Gibbs states and Markov random fields. *Advances of Applied Probability*, **5**, 242–61.

Richardson, T. S. (1996). *Models of feedback: interpretation and discovery.* PhD thesis, Carnegie-Mellon University.

Richardson, T. S. (1998). Chain graphs and symmetric associations. In *Learning in Graphical Models*, (ed. M. Jordan), pp. 231–60. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Richardson, T. S. (2001). Chain graphs which are maximal ancestral graphs are recursive causal graphs. Technical Report 387, Department of Statistics, University of Washington, Seattle.

Richardson, T. S. and Spirtes, P. (2000). Ancestral graph Markov models. Technical Report 375, Department of Statistics, University of Washington, Seattle.

Ripley, B. (1981). *Spatial Statistics.* John Wiley and Sons, New York.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximation. *Bernoulli*, **2**, 341–64.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods — application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**, 1393–512.

Speed, T. P. (1979). A note on nearest-neighbour Gibbs and Markov distributions over graphs. *Sankhya Ser. A*, **41**, 184–97.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Science*, **8**, 219–83.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Uncertainty in Artificial Intelligence: Proceedings of the* 11^{th} *Conference*, (ed. P. Besnard and S. Hanks), pp. 491–8. Morgan Kaufmann, San Francisco.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causality, Prediction and Search.* Springer-Verlag, New York.

Spirtes, P., Meek, C., and Richardson, T. S. (1995). Causal inference in the presence of latent variables and selection bias. In *Uncertainty in Artificial Intelligence: Proceedings of the* 11th *Conference*, (ed. P. Besnard and S. Hanks), pp. 403–10. Morgan Kaufmann, San Francisco.

Spirtes, P. and Richardson, T. S. (1997). A polynomial-time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Preliminary papers of the Sixth International Workshop on AI and Statistics, January 4-7, Fort Lauderdale, Florida*, (ed. D. Madigan and P. Smyth), pp. 489–501.

Spirtes, P., Richardson, T. S., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural modelling tool. *Sociological Methods and Research*, **27**, 182–225.

Spitzer, F. (1971). *Random Fields and Interacting Particle Systems.* Mathematical Association of America, Washington, DC. Notes on lectures given at the 1971 MAA Summer Session, Williams College, Williamstown, MA.

Strotz, R. H. and Wold, H. O. A. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, **28**, 417–27.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, (ed. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer), pp. 255–70. North-Holland, Amsterdam.

Wermuth, N. (1992). Block-recursive regression equations (with discussion). *Revista Brasileira de Probabilidade e Estatistica*, **6**, 1–56.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *Journal of the Royal Statistical Society, Series B*, **52**, 21–72.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* John Wiley and Sons, Chichester, United Kingdom.

Wold, H. O. A. (1953). *Demand Analysis.* John Wiley and Sons, New York. In association with L. Juréen.

Wold, H. O. A. (1954). Causality and econometrics. *Econometrica*, **22**, 162–77.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**, 557–85.