J. R. Statist. Soc. B (2002)
64, *Part* 3, *pp.* 321–361

Chain graph models and their causal interpretations

Steffen L. Lauritzen

Aalborg University, Denmark

and Thomas S. Richardson

University of Washington, Seattle, USA

[*Read before* The Royal Statistical Society *at a meeting organized by* the Research Section *on Wednesday, December 12th, 2001*, Professor D. Firth *in the Chair*]

Summary. Chain graphs are a natural generalization of directed acyclic graphs and undirected graphs. However, the apparent simplicity of chain graphs belies the subtlety of the conditional independence hypotheses that they represent. There are many 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 from the equilibrium distributions of dynamic models with feed-back. These dynamic interpretations lead to a simple theory of intervention, extending the theory developed for directed acyclic graphs. Finally, we contrast chain graph models under this interpretation with simultaneous equation models which have traditionally been used to model feed-back in econometrics.

Keywords: Causal model; Chain graph; Feed-back system; Gibbs sampler; Intervention theory; Structural equation model

1. Introduction

The use of directed acyclic graphs (DAGs) simultaneously to represent causal hypotheses and to encode independence and conditional independence constraints associated with those hypotheses may be traced back to the pioneering work of Wright (1921). More recently, DAGs have proved 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 (Spirtes *et al.*, 1993; Pearl, 1993, 1995, 2000; Lauritzen, 2001).

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, 1974a; 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

Address for correspondence: Steffen L. Lauritzen, Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9200 Aalborg, Denmark.
E-mail: steffen@math.auc.dk

 1369–7412/02/64321

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 for example 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 many 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 from equilibrium distributions of dynamic models with feed-back over time. Here again we shall 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 that has been developed for DAGs. Finally, we contrast chain graph models with simultaneous equation models which have traditionally been used to model feed-back in econometrics.

2. Basic graphical concepts and notation

In this paper we consider graphs containing both directed ('$\rightarrow$') and undirected ('—') edges and largely use the terminology of Lauritzen (1996), where the reader can also find further details. Below we briefly list some of the most central concepts used in this paper.

A *partially directed cycle* in a graph $\mathcal{G}$ is a sequence of n distinct vertices $v_1, \ldots, v_n$ ($n \geqslant 3$), and $v_{n+1} \equiv v_1$, such that

(a) $\forall\, i\ (1 \leqslant i \leqslant n)$ either v_i—v_{i+1} or $v_i \leftarrow v_{i+1}$, and
(b) $\exists\, j\ (1 \leqslant j \leqslant n)$ such that $v_j \leftarrow v_{j+1}$.

A subset C of vertices is *complete* if there is an edge (directed or undirected) between every pair of vertices in C. A maximal complete subset is a *clique*.

If $u \rightarrow v$, the vertex u is a *parent* of v and, if u—v, u is a *neighbour* of v. The set of parents of v is denoted $\mathrm{pa}(v)$ and the set of neighbours is denoted $\mathrm{ne}(v)$. The set $\mathrm{bd}(v) = \mathrm{pa}(v) \cup \mathrm{ne}(v)$ is the *boundary* of v. For a subset $A \subseteq V$, we let

$$\mathrm{pa}(A) = \cup_{v \in A}\, \mathrm{pa}(v) \backslash A.$$

A *chain graph* is a graph in which there are no partially directed cycles. A chain graph in which there are no undirected edges is a *DAG*.

The *chain components* $\mathcal{T}$ of a chain graph are the connected components of the undirected graph obtained by removing all directed edges from the chain graph. In a DAG, all chain components are singletons.

For $A \subseteq V$, $\mathcal{G}_A$ denotes the subgraph which has A as vertex set and all edges inherited from $\mathcal{G}$. Such a subgraph is said to be *induced* by A. A *minimal complex* in a chain graph is an induced subgraph of the form

$$a \rightarrow v_1 \text{—} \cdots \qquad\qquad \cdots \text{—} v_r \leftarrow b.$$

Minimal complexes play a fundamental role for the chain graph Markov property, to be further described below.

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 paper is concerned with the chain graph Markov property defined in Lauritzen and Wermuth (1984, 1989) and Frydenberg (1990). There have been several alternative suggestions for associating a Markov property with a chain graph (Cox and Wermuth, 1993; Andersson *et al*., 1996, 2001), 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 again 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) in the form

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

Here and in the following, x_A denotes a configuration $(x_v)_{v \in A}$ of a subset of variables $A \subseteq V$.

The chain graph Markov property manifests itself through an outer factorization

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

where each factor further factorizes according to the graph as

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

Here $\mathcal{A}(\tau)$ are the complete sets in the undirected graph $(\mathcal{G}_{\tau \cup \mathrm{pa}(\tau)})^m$, obtained from the subgraph $\mathcal{G}_{\tau \cup \mathrm{pa}(\tau)}$ by 'moralization' (Lauritzen (1996), page 7), i.e. adding edges between unconnected parents of τ and ignoring directions on remaining edges. The factor Z is a normalizer

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

Note that the outer factorization (2) may be viewed as a DAG with vertices representing the multivariate random variables X_τ for $\tau \in \mathcal{T}$. Andersson *et al*. (1996) referred 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.

For subsets of variables A, B and S, the expression $A \perp\!\!\!\perp B | S$ denotes that the variables in A are conditionally independent of those in B, given the values of the variables in S (Dawid, 1979). We use the notation $A \not\perp\!\!\!\perp B | S$ to mean that the conditional independence of A and B given S is not a consequence of the global Markov property, implying that the conditional independence will fail for some (but not all) probability measures which factorize (Studený and Bouckaert, 1998).

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, proved in full generality by Andersson *et al.* (1997).

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.

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

4. Causal interpretation of directed acyclic graph 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 *effects of interventions* on associated distributions.

4.1. Conditioning by observation or intervention

We initially emphasize the distinction between different types of conditioning operations, each of which modifies a given probability distribution. Conditional densities are usually calculated as

$$f(y|x) = f(y|X = x) = f(y, x)/f(x).$$

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

In general this is not the way that 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 the two types of conditioning, as indicated below:

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

Other researchers have used expressions such as $P(Y_x = y)$, $P_{\mathrm{man}(x)}(y)$, $\mathrm{set}(X = x)$, $X = \hat{x}$ or $\mathrm{do}(X = x)$ to denote intervention conditioning (Neyman, 1923; Rubin, 1974; Spirtes *et al.*, 1993; Pearl, 1993, 1995, 2000).

Generally, the two quantities will be different,

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

and the quantity on the left-hand side cannot be calculated from the density alone, without additional assumptions. The difference has often mistakenly been ignored in statistical literature although there are examples, where the distinction is very clearly made; see for example Box (1966) or Cox (1984).

Below we shall give a precise causal interpretation of a DAG. This will imply that in the first graph below

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

4.2. Data-generating process for directed acyclic graph models

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

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

where the assignments must be carried out sequentially in a well ordering of the DAG $\mathcal{D}$, 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. For any given probability distribution, there is a multitude of choices for g_v and U_v in the generating process. Deriving results on the basis of this representation should therefore be made with extreme caution, to avoid undue dependence on the specific choice made (Dawid, 2000).

This assignment system can be seen as a general structural equation model (SEM) as invented in the context of genetics (Wright, 1921), and exploited in economics (Haavelmo, 1943; Wold, 1954) and social sciences (Goldberger, 1972). SEMs were also used as the main justification and motivation for studying directed Markov models in Kiiveri *et al*. (1984) and Kiiveri and Speed (1982). We shall return to these models in Section 7.

It is appropriate to think of a data-generating process as a 'computer program', well ordering the elements of V as in expression (4) so that $V = 1, \ldots, p$ and writing

for $i = 1, \ldots, p$;
  $\varepsilon \leftarrow$ runif;
  $x_i \leftarrow h_i(x_{\mathrm{pa}(i)}, \varepsilon)$;
return x;

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

It is an important aspect of SEMs that they also specify the way in which intervention is to be modelled. 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 modelled by replacing the corresponding line in expression (4) or the equivalent computer program with the assignment described by the intervention. We refer to this type of intervention as *intervention by replacement*.

4.3. Causal directed acyclic graphs

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

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

For $A = \varnothing$ 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 way of deriving different probability measures, each representing the probability law associated with a specific intervention.

We shall refer to equation (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 DAG $\mathcal{D}$ 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. Note in particular that the intervention distributions for variables in V are indeed independent of the particular choices of g_v and U_v.

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 for example 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 SEMs 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 $Z \subseteq V \backslash \{x, y\}$ so that $x \perp\!\!\!\perp y | Z$. 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 to consider graphs in which all edges are directed, since to do so would rule out the pos-sibility 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 and y. For example

(a) the association may have arisen due to the presence of
    (i) an unmeasured confounding variable,
    (ii) some artefact of the way that the sample was selected or
    (iii) a feed-back relationship, or
(b) 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 (a) and (b): in situation (b), additional knowledge might justify including an edge $x \rightarrow y$, whereas this is not so with (a). In philosophical terms, reasons under (a) would be described as *ontological*; those under (b) as *epistemological*.

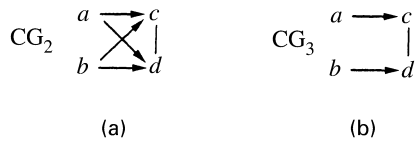

Fig. 1. Two examples of chain graphs in which c and d are joint responses to a and b

Although the original papers on chain graphs are clear that directed edges are to be interpreted as (in some sense) causal, whereas 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 Fig. 1(a) 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 and d as being 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 such as graph CG_3 in Fig. 1(b) restricts $f(c, d|a, b)$. Under the chain graph Markov property, graph 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 cannot be interpreted in any of the ways listed above other than feed-back.

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

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

corresponding to an independence structure that is different from that of graph 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:

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

Consequently, neither of these generating processes explains graph CG_3 of Fig. 1(b).

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

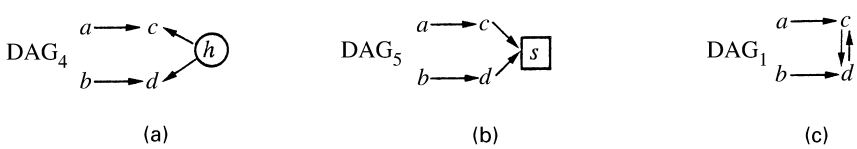

Fig. 2. (a), (b) Generating processes in which c and d are on an equal footing, that do not give rise to the conditional independence model given by graph CG_3 under the standard Markov property; (c) directed cyclic graph, corresponding to a non-recursive linear SEM, again not Markov equivalent to graph CG_3

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

which again does not correspond to graph 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, graph 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

We 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 graph CG_3. This was pointed out by Richardson (1998), who showed that all conditional independence structures which can be obtained by such marginalization and conditioning from a DAG satisfy a property of *between separation* (theorem 1 of Richardson (1998)), whereas graph 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 that are obtainable from some DAG by marginalization and conditioning, as stated more accurately in the following proposition.

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 DAG $\mathcal{D}$ over $V \cup L \cup S$, if and only if $\mathcal{K}$ is Markov equivalent to a recursive causal graph.

5.2. Chain graphs as unions of directed acyclic graph 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 DAG is elicited from an expert (Jensen, 1996; Spiegelhalter *et al*., 1993).

If $\mathcal{D}_1$ and $\mathcal{D}_2$ are two DAGs with the same set of adjacencies but, for some pair(s) 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 replacing common edges of different directions with undirected edges may contain edges of both types. Note that the edge set of $\mathcal{D}_{1\cup 2}$ is the union of the edge sets of $\mathcal{D}_1$ and $\mathcal{D}_2$; see Lauritzen (1996), page 4.

However, as exemplified in Fig. 3, a graph produced by taking unions of DAGs will only be a chain graph in quite special cases. Two such cases are

(a) when $\mathcal{D}_1$ and $\mathcal{D}_2$ have the same adjacencies but differ over the orientation of a single edge only and
(b) when a graph is formed by taking the union of all DAGs which are Markov equivalent to a given DAG (Andersson *et al*., 1997).

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 $\mathcal{D}_2$, which

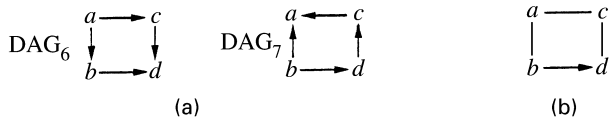

Fig. 3. (a) Two DAGs with the same sets of adjacencies and (b) the graph formed from (a) by representing edges of different direction with undirected edges

would be the model obtained by assuming that the direction of certain edges is unknown. 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.

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) showed 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 are vertices v_1, v_2, $\alpha \in V$ such that v_1 and v_2 are not adjacent, and $v_1 \to \alpha \leftarrow v_2$ in one graph, but in the other

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

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—α or α—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 $v_1 \perp\!\!\!\perp v_2 | T$. However, it is easy to construct a distribution in $M(\mathcal{D}_1)$ in which $v_1 \not\perp\!\!\!\perp v_2 | T$ for any set T containing α. Suppose for example that all variables take states 0 and 1. Then let

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

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, whereas 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

(a) variables may be divided into *risk factors*, *diseases* and *symptoms*,
(b) in a longitudinal study variables may be grouped according to time and

(c) 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, when attention is restricted to chain graph models.

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 Fig. 4(b) and that, in fact, the simple causal graph DAG_1 in Fig. 4(a) 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 graph DAG_1 is saturated, as shown in Fig. 4(c). Thus a search for a chain graph model that is 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 prespecified blocking may preclude finding the most parsimonious model. It is also simple to see that if a, x and y had 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. For illustration we only discuss 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 and 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, whereas $v \prec w$ indicates that the component containing v precedes the component containing w in the ordering. Note that for cases 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.

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), page 55) that in cases 1–5 $x_1 \perp\!\!\!\perp x_2 | z$, whereas in case 6 $x_1 \perp\!\!\!\perp x_2$.

We shall show by example that for each of the orderings specified in Table 1 there are 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.

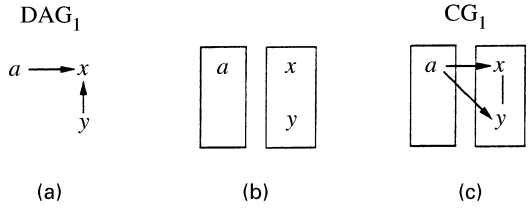

Fig. 4. Restricting to chain graph models in keeping with a block ordering may lead to less parsimonious models: (a) DAG_1, the generating process; (b) a block ordering $\{a\} \prec \{x, y\}$; (c) CG_1, the minimal chain graph model for $\{a, x, y\}$, compatible with the ordering which contains the model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

Table 1. Chain graphs with three vertices and two edges

Case	*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	$x_1 \prec z \succ x_2$	$x_1 \rightarrow z \leftarrow x_2$	$x_1 \perp\!\!\!\perp x_2$

For cases 1–5 of Table 1 consider graph DAG_3 shown in Fig. 5(a), 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 the second column 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 graph DAG_4 shown in Fig. 5(b). Whereas in DAG_3 we marginalized h_1 and h_2, here we consider $P(x_1, x_2, z|s)$. A simple interpretation of s is that it represents a *selection* variable, which takes the same value for all units in the subpopulation being modelled. See Cox and Wermuth (1996), page 44, Cooper (1995), Spirtes *et al.* (1995), Spirtes and Richardson (1997) and Lauritzen (1999) for further discussion. In the conditional distribution $P(x_1, x_2, z|s)$ it holds that $x_1 \perp\!\!\!\perp x_2|z$, rather than marginal independence, which is implied by the chain graph under this ordering.

In fact it may be shown that any DAG that is compatible with the ordering given by case 6 in which no pairwise marginal independence holds among the variables x_1, x_2 and z, but which satisfies $x_1 \perp\!\!\!\perp x_2|z$, will contain variables that are conditioned on; marginalization alone is not sufficient. This may explain why it is often inferred that conditional independence given z is incompatible with $x_1 \prec z \succ x_2$. See for example Mohamed *et al.* (1998), page 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 that is compatible with

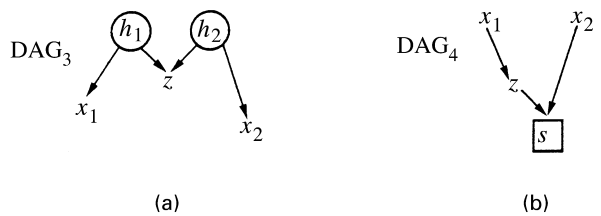

Fig. 5. Examples showing that, when latent or selection variables may be present, ordering of variables does not imply the conditional independence relations given by a chain graph: in graph DAG_3 h_1 and h_2 are hidden variables; in graph DAG_4, s is a selection variable that has been conditioned on

those blocks; additional detailed substantive arguments, ruling out (or hypothesizing) the absence of confounders, are always required.

We conclude this section by making some further points.

(a) 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 $x_1 \perp\!\!\!\perp x_2 | z$ and $x_1 \perp\!\!\!\perp x_2$ we consider $x_1 \perp\!\!\!\perp x_2 | W$, with $z \in W$ and $z \notin W$ respectively.

(b) The problems which 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.

(c) There are independence structures arising from DAGs with hidden variables that cannot be represented by any chain graph model. Fig. 2(a) is an example. Wermuth *et al.* (1994, 1999), Koster (1999, 2000) and Richardson and Spirtes (2000) have provided graphical representations 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?

(d) Often it appears that resistance to consideration of models that violate blocking follows from a naïve causal interpretation of the resulting graph. Thus for instance, if graph DAG_3 in Fig. 5(a) is the generating process, then the independence structure can be represented by the chain graph $x_1 \rightarrow z \leftarrow x_2$. 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.

(e) In some cases, more principled objections to consideration of a less restricted class of chain graphs may be adduced: computational issues may be 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 are 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. Feed-back 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 might suggest that a general data-generating process for chain graph models would involve infinite processes converging to some type of equilibrium.

In this section we present some alternative equilibrium data-generating processes with feed-back 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 of $\mathcal{G}$. Such graphical models originate 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.

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 be written in an idealized form 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 $f(y_i|x_{-i})$;
return x.

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

The expressions 'until equilibrium' and 'return x' must 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 'snapshot' in time of the configurations of the variables.

The system involves feed-back in the sense that the value of X_i for any $i \in V$ has been dynamically affected by all 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:
$v \leftarrow \text{rand}(V)$;
$x_v \leftarrow y_v$ with probability $f(y_v|x_{-v})$;
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

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

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

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

$$X(t+\mathrm{d}t) = X(t) + \tfrac{1}{2}\mathrm{grad}(\log[f\{X(t)\}])\,\mathrm{d}t + \mathrm{d}W(t) \tag{7}$$

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 owing 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 0 and a non-singular covariance matrix Σ with inverse $K = \Sigma^{-1}$. The distribution satisfies the Markov property of an undirected graph if and only if we have

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

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

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

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, expression (8) implies that the sum above only extends over the neighbours of v. We shall write this dynamic scheme as

$$X(t+1) \overset{G}{\leftarrow} A * X(t) + \varepsilon(t+1) \tag{9}$$

where A is the matrix of coefficients. The special assignment symbol and asterisk indicate 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), such a matrix 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 diagonal elements 0 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 $\varepsilon_v(t)$. If the variables are scaled to have innovation variances 1, 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 diagonal elements 0 and that $I - A$ be symmetric and positive definite (Besag, 1975; Ripley, 1981). The covariance matrix of the equilibrium distribution is then given by $\Sigma = (I - A)^{-1}$.

If A does not satisfy these conditions, the behaviour of the updating scheme will typically depend on the ordering of the variables and several patterns of behaviour are possible; see Appendix A.

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

$$X(t + \mathrm{d}t) = X(t) - \tfrac{1}{2}KX(t)\,\mathrm{d}t + \mathrm{d}W(t). \tag{11}$$

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

$$X(t + \mathrm{d}t) = X(t) + CX(t)\,\mathrm{d}t + \mathrm{d}Z(t), \tag{12}$$

where $Z(t)$ is Brownian motion with covariance matrix $\mathbf{V}\{\mathrm{d}Z(t)\} = \Lambda$; 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 0 and covariance matrix equal to the unique solution of the matrix equation

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

Clearly there are many more choices for C and Λ leading to $\Sigma = K^{-1}$ than $C = -K/2$ used in the Langevin diffusion (11). Proposition 5 below shows that this choice has a distinguished intervention property.

6.2. Intervention in undirected graphs

Each of the dynamic schemes described above corresponds in a natural way to an intervention model. 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 modelled 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

$$X_B(t + \mathrm{d}t) = X_B(t) + \tfrac{1}{2}\mathrm{grad}(\log[f\{X_B(t), x_A^*\}])\,\mathrm{d}t + \mathrm{d}W_B(t). \tag{15}$$

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 condition (14) also holds in this case.

If we consider a more general dynamic regime such as the diffusion (12) this may no longer be true. Indeed we have the following result in the Gaussian case.

Proposition 5. Let P be the equilibrium distribution of the diffusion process (12) with $\Lambda = I$. If intervention is made by replacement, then

$$P(X_B = x_B | X_A \leftarrow x_A^*) = P(X_B = x_B | X_A = x_A^*)$$

if and only if C is symmetric and negative definite. It then holds that $C = -\Sigma^{-1}/2$, where Σ is the covariance matrix of the equilibrium distribution.

Proof. If C is symmetric it is a stability matrix if and only if it is negative definite. Then the unique solution to equation (13) is clearly $\Sigma = -C^{-1}/2$. Thus equation (12) is the Langevin diffusion and the intervention formula (14) holds.

Next, assume that formula (14) holds. The effect of an intervention under this diffusion leads to

$$X_B(t + \mathrm{d}t) = X_B(t) + C_{BB}\,X_B(t)\,\mathrm{d}t + C_{BA}x_A^*\,\mathrm{d}t + \mathrm{d}Z_B(t),$$

where the matrix C has been partitioned into appropriate blocks. The equilibrium distribution of the intervention diffusion has expectation equal to

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

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

$$I + C_{BB}\Omega_{BB} + \Omega_{BB}C_{BB}^{\mathrm{T}} = 0.$$

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

$$C_{BB}^{-1}C_{BA} = K_{BB}^{-1}K_{BA} \tag{16}$$

and

$$I + C_{BB}K_{BB}^{-1} + K_{BB}^{-1}C_{BB}^{\mathrm{T}} = 0. \tag{17}$$

From the special case where $B = \{v\}$ is a singleton, we obtain from equation (17)

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

and inserting this into equation (16) 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. In particular this implies that C is symmetric and negative definite.

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.

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:

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

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 $x_{\mathrm{pa}(\tau)}$;
  $x_\tau \leftarrow x_\tau^0$;
  repeat until equilibrium:
    $v \leftarrow \mathrm{rand}(\tau)$;
    $x_v \leftarrow y_v$ with probability $f(y_v | x_{\tau \setminus \{v\}}, x_{\mathrm{pa}(\tau)})$;
return x_τ

and similarly in its systematic form. 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.

This 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 $\varepsilon^\tau = (\varepsilon^1, \varepsilon^2, \ldots)$ denote a 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 $(x_{\mathrm{pa}(\tau)}, \varepsilon^\tau)$;
  $x_\tau \leftarrow x_\tau^0$;
  $n \leftarrow 0$;
  repeat until equilibrium:
    $v \leftarrow$ rand(τ);
    $n \leftarrow n + 1$;
    $x_v \leftarrow h_v^\tau(x_{\tau\setminus\{v\}}, x_{\mathrm{pa}(\tau)}, \varepsilon^n)$;
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 $f(y_v|x_{\tau\setminus\{v\}}, x_{\mathrm{pa}(\tau)})$, 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 obtain that

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

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

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

Proposition 6. 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 assignment 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 $f(x_\tau|x_{\mathrm{pa}(\tau)})$ 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 equation (2) as desired. □

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

Since equilibrium may not be attained in finite time, 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 the 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 in the data-generating processes of Section 6.3 by replacement in each chain component as described in Section 6.2, it follows as in the directed case that this leads to the formula

$$p(x||x_A) = \prod_{\tau \in \mathcal{T}} p(x_{\tau \backslash A} | x_{\mathrm{pa}(\tau)}, x_{\tau \cap A}). \tag{18}$$

This specializes to the intervention formula (5) in the fully directed case and Bayes's formula in the undirected case: in the fully directed case, all chain components are singletons, so either $\tau \backslash A$ or $\tau \cap B$ are empty; in the undirected case $\mathrm{pa}(\tau)$ are all empty. The formula also conforms with the calculus of decision networks based on chain graphs as discussed in Cowell *et al.* (1999), where interventions are then described by decision nodes. Since

$$p(x_{\tau \backslash A} | x_{\mathrm{pa}(\tau)}, x_{\tau \cap A}) = \tilde{Z}^{-1}(x_{\mathrm{pa}(\tau)}, x_{\tau \cap A}) \prod_{C \in \mathcal{A}(\tau)} \phi_C(x_C),$$

where $\tilde{Z}$ is a normalizer as before, an alternative argument for formula (18) may be based on the assumption that the potentials $\psi_C = \log(\phi_C)$ 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.

6.5. Equilibrium dynamics and infinite directed acyclic graphs

It is illuminating to think of the equilibrium dynamics described in terms of infinite DAGs. If, for example, we consider the simple chain graph CG_3 in Fig. 1(b), 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, and then use these as input for an equilibrium process updating of c and d as indicated in Fig. 6. Using the global Markov property on the DAG in Fig. 6 yields

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

whereas in general

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

since b and c_i are common parents of d_i in the update scheme described.

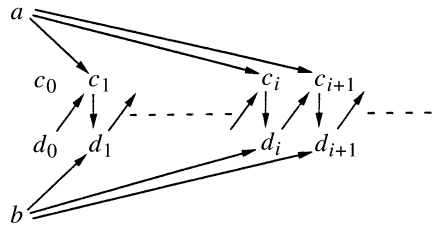

Fig. 6. Infinite DAG corresponding to a structural assignment system for chain graph CG_3 of Fig. 1 where c is updated before d in each inner loop

Thus taking a snapshot as

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

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

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 holds under Gibbs dynamics), then (X_{c_i}, X_{d_i}) and $(X_{c_i}, X_{d_{i-1}})$ have the same equilibrium distribution; see Appendix A. It therefore holds *in equilibrium*—and thus approximately for large i—that $c_i \perp\!\!\!\perp b | \{d_i, a\}$, provided that such update distributions are used.

7. Linear structural equation models

7.1. Basic terminology

In a linear SEM, variables are conventionally divided into two disjoint sets: substantive variables and error variables (Bollen, 1989). A further distinction between 'exogenous' and 'endogenous' substantive variables is sometimes made; we have not done so as it is not relevant to our discussion.

A unique error term $\boldsymbol{\varepsilon}_v$ is associated with each substantive variable X_v, $v \in V$. 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

$$X = \Gamma X + \boldsymbol{\varepsilon}, \tag{19}$$

where $\gamma_{vv} = 0$. In any given structural model some off-diagonal entries in Γ may also be fixed at 0, 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 0, an 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.

The term 'equation' is really misplaced, and it seems more appropriate to use the term 'structural assignment model' and to write expression (19) as

$$X \leftarrow \Gamma X + \varepsilon.$$

If Γ is lower triangular and has 0 in the diagonals, this expression can be given an unam-

biguous meaning by making the assignment sequentially by row, but in general it is not obvious which meaning to attribute to such an assignment symbol.

In the traditional interpretation of an SEM a multivariate normal distribution over the error terms is specified as for example $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Delta)$. In any particular model, some off-diagonal (δ_{ij}) entries in Δ may be allowed 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 traditional interpretation of an SEM determines a joint distribution over the substantive variables by solving equations (19) to obtain the *reduced form* equations

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

yielding

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

Much controversy and confusion in the literature is due to treating the assignment systems as equation systems in this way and uncritically moving variables between the left-hand and the right-hand side of expression (19). This can make a radical difference, in particular when the effects of interventions are considered. See for example Pearl (1998) and Spirtes *et al.* (1998) for a detailed discussion of these and other issues concerning SEMs.

The distribution obtained in the traditional way 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 can naturally be associated with the directed graph of Fig. 2(c) with a relabelling of the vertices as $(a, b, c, d) = (1, 2, 3, 4)$:

$$x_1 = \varepsilon_1,$$

$$x_2 = \varepsilon_2,$$

$$x_3 = \gamma_{31}x_1 + \gamma_{34}x_4 + \varepsilon_3,$$

$$x_4 = \gamma_{42}x_2 + \gamma_{43}x_3 + \varepsilon_4,$$

$$\Delta = \begin{pmatrix} \delta_{11} & 0 & 0 & 0 \\ 0 & \delta_{22} & 0 & 0 \\ 0 & 0 & \delta_{33} & 0 \\ 0 & 0 & 0 & \delta_{44} \end{pmatrix}.$$

Fisher (1970) presented a dynamic process whose time average gives the distribution described by a linear non-recursive SEM. Here the system is occasionally subjected to random exogenous 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.

This equilibrium interpretation can thus be seen as being deterministic, but with random boundary conditions. In the next section we discuss an interpretation of non-recursive structural equations in terms of stochastic equilibrium.

As mentioned, using the intervention interpretation of structural equations given by Strotz and Wold (1960) leads here to an intervention distribution which is different from those earlier described. Indeed, if in the example given we intervene as $X_4 \leftarrow x_4^*$ we obtain the

recursive SEM

$$\left.\begin{aligned} x_1 &= \varepsilon_1, \\ x_2 &= \varepsilon_2, \\ x_3 &= \gamma_{31}x_1 + \gamma_{34}x_4^* + \varepsilon_3, \\ \Delta &= \begin{pmatrix} \delta_{11} & 0 & 0 \\ 0 & \delta_{22} & 0 \\ 0 & 0 & \delta_{33} \end{pmatrix}. \end{aligned}\right\} \tag{20}$$

7.2. Chain graph models for structural equations

The chain graph models and corresponding generating processes can in some cases give an alternative interpretation of a structural equation system with coefficient matrix Γ.

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

The graph of an 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 it is, the dynamic schemes discussed in Section 6 could be used to give an alternative interpretation of an SEM with feed-back.

Then, in each chain component of the graph, the structural equations are interpreted as conditional autoregressions. More accurately, the chain components are first ordered in a sequence that is compatible with the chain graph and then each part of the assignment system is interpreted through Gibbs updating as

$$X_\tau(t+1) \overset{G}{\leftarrow} \Gamma_\tau * X_\tau(t) + \Gamma_{\tau,\mathrm{pa}(\tau)} * x_{\mathrm{pa}(\tau)} + \varepsilon_\tau(t+1)$$

where the subscripted matrices are appropriate submatrices of Γ and the asterisk denotes that the update is to be made sequentially by row.

As mentioned in Section 6.1.5, such a specification does not always correspond to a well-defined distribution. The system should satisfy

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

Thus there is only a single free parameter to describe the relation between two variables instead of two as in a conventional SEM. In addition—if we again assume that the variables have been scaled to have error variances 1—the submatrices $I_\tau - \Gamma_\tau$ induced by the corresponding chain component would have to be positive definite. In the example considered, these conditions would amount to

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

The first condition ensures balance whereas the second condition ensures stability of the dynamic system. Arnold *et al.* (1999) investigated this bivariate case in detail.

Thus, non-recursive SEMs 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 SEM (20), but now it is obtained from the joint distribution by the intervention

formula (18). The joint distribution is different under the chain graph interpretation of the SEM, for which expression (19) would not lead to the distribution (20).

Ord (1976) also suggested the use of the CAR interpretation for simultaneous equation models in economics, whereas Wermuth (1992) suggested quite a different chain graph representation of simultaneous equations with other special restrictions on the parameters; see Lauritzen (1996), pages 154–155.

8. Discussion

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

8.1. Causal directed acyclic graphs versus causal chain graphs

There is a large body of work which takes as its starting-point the assumption that the variables in the population of interest were generated by a causal DAG as described in Section 4, possibly with some variables unobserved. The considerations in Section 6 indicate that in some circumstances this assumption may be unduly restrictive: if feed-back is present then the model for the equilibrium distributions of 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 *versus* non-recursive SEMs.

8.2. Undirected edges and causal underdetermination

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.

(a) 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 feed-back.
(b) Chain graphs do not in general represent the independence structures that arise from DAGs with hidden variables. For this, other types of graph are required.
(c) 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; refraining from assigning a direction to an edge may amount to making a definite commitment to a particular causal hypothesis. Further, there are alternative 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 that are compatible with a prespecified ordering will often be incompatible with finding the most parsimonious model. This seems undesirable.

(a) If the primary goal of the analysis is prediction (of the joint distribution) then parsimonious models are often preferable.

(b) If explanation is the goal then a less parsimonious model—which will include 'extra' edges—may often be misleading; see Fig. 4.

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' (Spirtes *et al*., 1993) or 'lack of stability' (Pearl, 2000)—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. However, if we have good reason to believe that parametric cancellation is present, then this might argue against attempting to model the independence structure to understand the generating process.

The alternative of directly modelling the conditional independence structure without assuming that it arises from a generating process appears intractable; even for only four variables there are 18 300 such structures for discrete distributions; see Matúš (1999) and references therein.

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* a model search has been completed 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, 2001). Hence, in general, different statistical models may be associated with the same chain graph. For example, with this alternative interpretation the graph CG_3 in Fig. 1(b) encodes 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 graph DAG_4 in Fig. 2(a). 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 feed-back system. This naturally raises the question which generating processes correspond to chain graphs under this alternative Markov property. Cox and Wermuth (1993) discussed other possible ways of encoding conditional independence relations using chain graphs, for which the same question may arise.

8.5. Conclusion

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' (page 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.

Acknowledgements

This research was supported in part by the Danish Research Councils through their programme in information technology under the Danish Informatics Network in Agricultural Sciences project and the US National Science Foundation Division of Mathematical Sciences (grant DMS-9972008). In addition the authors gratefully acknowledge inspiration and support from the European Science Foundation scientific programme on highly structured stochastic systems and the Isaac Newton Institute where the second author was a Rosenbaum Fellow from July to December 1997.

Appendix A: Limiting behaviour of Gibbs dynamics

In the following we let $q = (q_v)_{v \in V}$ denote a family of conditional specifications, i.e. $q_v(\cdot|x_{-v})$ denotes for all $x_{-v} \in \mathcal{X}_{V\setminus\{v\}}$ a probability distribution over $\mathcal{X}_v$. For simplicity we assume that the support of $q_v(\cdot|x_{-v})$ is equal to $\mathcal{X}_v$ for all $v \in V$, i.e. that

$$q_v(A|x_{-v}) > 0 \qquad \text{for all open sets } A \subseteq \mathcal{X}_v. \tag{22}$$

We say that q is *consistent* if there is a probability measure μ on $\mathcal{X}$ such that, for all $v \in V$, q_v is a version of the conditional distribution with respect to μ of X_v, given $X_{-v} = x_{-v}$, i.e. if there is a μ satisfying the equation

$$\mu(A) = \int_{x_{-v}} q_v(A|x_{-v})\,\mu_{-v}(\mathrm{d}x_{-v}), \qquad \text{for all } v \in V. \tag{23}$$

If we introduce the transition kernel Q_v

$$Q_v(A|x) = q_v(A|x_{-v}),$$

we may rewrite equation (23) in a shorter form:

$$\mu = \mu Q_v, \qquad \text{for all } v \in V.$$

If q is consistent, we know that the Gibbs sampler forms a Markov chain which converges to the uniquely determined equilibrium distribution. We shall briefly discuss the possible behaviour of the systematic Gibbs sampler in cases where q is not necessarily consistent.

So consider V numbered as $V = \{1, \ldots, p\}$ and define for each permutation $\pi \in S(p)$ the transition kernel

$$P(\pi) = Q_{\pi(1)} Q_{\pi(2)} \ldots Q_{\pi(p)}.$$

Then $P(\pi)$ is the transition kernel of the Markov chain formed by the systematic Gibbs sampler using q as its update distribution, and updating the sites in the order determined by π. Condition (22) ensures that this Markov chain is irreducible and aperiodic. We then have the following results.

Lemma 1. Let e be the identity permutation. Then $P(e)$ has an invariant distribution if and only if $P(\sigma)$ has an invariant distribution for all cyclic permutations σ.

Proof. First we show that if μ is an invariant distribution for $P(e)$ then $\mu Q_1 \ldots Q_{i-1}$ is invariant for $P(\sigma_i)$, where $\sigma_i = (i, \ldots, p, 1, \ldots, i-1)$. This follows from the calculation

$$\mu Q_1 \ldots Q_{i-1} P(\sigma_i) = \mu P(e) Q_1 \ldots Q_{i-1} = \mu Q_1 \ldots Q_{i-1}.$$

The converse follows by renumbering V. □

Consequently we obtain the following proposition.

Proposition 7. The following conditions are equivalent for a probability measure μ on $\mathcal{X}$:

(a) $\mu P(\pi) = \mu$ for all $\pi \in S(p)$;
(b) $\mu P(\sigma) = \mu$ for all $\sigma \in S(p)$ with σ cyclic;
(c) $\mu = \mu Q_i$ for all $i \in V$.

Proof. We show that (a) implies (b) implies (c) implies (a). The implication (a)–(b) is trivial. If condition (b) holds, and $i \in V$, we obtain

$$\mu Q_i P(\sigma_{i+1}) = \mu P(\sigma_i) Q_i = \mu Q_i.$$

Thus μQ_i is an invariant distribution for $P(\sigma_i)$. As the invariant distribution is uniquely determined, we must have $\mu Q_i = \mu$ as required for condition (c).

That condition (c) implies (a) is easily shown by the repeated use of the relations $\mu Q_i = \mu$:

$$\mu P(\pi) = \mu Q_{\pi(1)} Q_{\pi(2)} \cdots Q_{\pi(p)} = \mu Q_{\pi(2)} \cdots Q_{\pi(p)} = \ldots = \mu Q_{\pi(p)} = \mu.$$

This completes the proof. □

Further, we have the following corollary.

Corollary 1. The specifications q are consistent if and only if, for all permutations $\pi \in S(p)$, $P(\pi)$ has an invariant distribution $\mu(\pi)$ which is independent of π.

Inconsistency of q might thus show up in two different ways. It may happen that $P(\pi)$ is transient, in which case there is no invariant distribution and the Gibbs sampler will drift away. Alternatively, it may exhibit stationary behaviour, but with a limiting distribution depending on the particular choice of ordering π in the sitewise updating. If the state space is finite, transient behaviour is not possible.

If the state space is infinite and $|V| \geqslant 3$, it may be that the Gibbs sampler converges to equilibrium for one permutation π but shows transient behaviour for another permutation π', provided that π and π' are not cyclically equivalent.

Stationary behaviour of the Gibbs sampler in the inconsistent case can be particularly dangerous in certain applications, as in most cases only a single ordering is chosen or the random updating scheme is used. The Gibbs sampler will then, without any warning signals, converge to a limiting distribution μ, but the specifications q will not be conditional distributions with respect to this μ and the results obtained may thus be misleading.

It has been suggested (Hofmann and Tresp, 1998; Hofmann, 2000; Heckerman *et al*., 2000) to use what Heckerman *et al*. (2000) termed the 'pseudo-Gibbs sampler' in any case, in particular when the distributions are expected to be almost consistent. This could, for example, be expected when the specifications q have been determined from empirical data. However, it would be desirable to have a more precise understanding of the general relation between the limiting distribution μ of a stationary pseudo-Gibbs sampler and the conditional specifications q.

References

Andersson, S. A., Madigan, D. and Perlman, M. D. (1996) An alternative Markov property for chain graphs. In *Proc. 12th Conf. Uncertainty in Artificial Intelligence* (eds F. V. Jensen and E. Horvitz), pp. 40–48. San Francisco: Morgan Kaufmann.

——— (1997) A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist*., **25**, 505–541.

——— (2001) Alternative Markov properties for chain graphs. *Scand. J. Statist*., **28**, 33–85.

Arnold, B., Castillo, E. and Sarabia, J. M. (1999) *Conditionally Specified Distributions*. New York: Springer.

Bentzel, R. and Hansen, B. (1954) On recursiveness and interdependency in economic models. *Rev. Econ. Stud*., **22**, 153–168.

Besag, J. (1974a) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc*. B, **36**, 302–339.

——— (1974b) On spatial-temporal models and Markov fields. In *Trans. 7th Prague Conf. Information Theory, Statistical Decision Functions and Random Processes*, pp. 47–55. Prague: Academia.

——— (1975) Statistical analysis of non-lattice data. *Statistician*, **24**, 179–195.

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

Box, G. E. P. (1966) Use and abuse of regression. *Technometrics*, **8**, 625–629.

Cooper, G. F. (1995) Causal discovery from data in the presence of selection bias. In *Preliminary Pap. 5th Int. Wrkshp AI and Statistics, Jan. 4th–7th, Fort Lauderdale* (ed. D. Fisher), pp. 140–150.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.
Cox, D. R. (1984) Design of experiments and regression. *J. R. Statist. Soc.* A, **147**, 306–315.
Cox, D. R. and Wermuth, N. (1993) Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.*, **8**, 204–218, 247–277.
——— (1996) *Multivariate Dependencies: Models, Analysis and Interpretation*. London: Chapman and Hall.
——— (2000) On the generation of the chordless four-cycle. *Biometrika*, **87**, 204–212.
Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980) Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.*, **8**, 522–539.
Dawid, A. P. (1979) Conditional independence in statistical theory (with discussion). *J. R. Statist. Soc.* B, **41**, 1–31.
——— (2000) Causal inference without counterfactuals. *J. Am. Statist. Ass.*, **95**, 407–448.
Edwards, D. and Kreiner, S. (1983) The analysis of contingency tables by graphical models. *Biometrika*, **70**, 553–562.
Fisher, F. M. (1970) A correspondence principle for simultaneous equation models. *Econometrica*, **38**, 73–92.
Frydenberg, M. (1990) The chain graph Markov property. *Scand. J. Statist.*, **17**, 333–353.
Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.
Gibbs, W. (1902) *Elementary Principles of Statistical Mechanics*. NewHaven: Yale University Press.
Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall.
Goldberger, A. S. (1972) Structural equation models in the social sciences. *Econometrica*, **40**, 979–1002.
Grenander, U. and Miller, M. I. (1994) Representations of knowledge in complex systems (with discussion). *J. R. Statist. Soc.* 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.
Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R. and Kadie, C. (2000) Dependency networks for inference, collaborative filtering, and data visualization. *J. Mach. Learn. Res.*, **1**, 49–75.
Hofmann, R. (2000) Inference in Markov blanket networks. *Technical Report FKI-235-00*. Technical University of Munich, Munich.
Hofmann, R. and Tresp, V. (1998) Non-linear Markov networks for continuous variables. In *Advances in Neural Information Processing Systems 10* (eds M. I. Jordan, M. J. Kearns and S. A. Solla), pp. 521–527. Cambridge: MIT Press.
Jensen, F. V. (1996) *An Introduction to Bayesian Networks*. London: University College London Press.
Kiiveri, H. and Speed, T. P. (1982) Structural analysis of multivariate data: a review. In *Sociological Methodology* (ed. S. Leinhardt). San Francisco: Jossey-Bass.
Kiiveri, H., Speed, T. P. and Carlin, J. B. (1984) Recursive causal models. *J. Aust. Math. Soc.* A, **36**, 30–52.
Koster, J. T. A. (1996) Markov properties of non-recursive causal models. *Ann. Statist.*, **24**, 2148–2177.
——— (1999) Linear structural equations and graphical models. *Lecture Notes*. Fields Institute, Toronto.
——— (2000) Marginalizing and conditioning in graphical models. *Technical Report*. Erasmus University, Rotterdam.
Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon.
——— (1999) Generating mixed hierarchical interaction models by selection. *Technical Report R-99-2021*. Department of Mathematical Sciences, University of Aalborg, Aalborg.
——— (2001) Causal inference from graphical models. In *Complex Stochastic Systems* (eds O. E. Barndorff-Nielsen, D. R. Cox and C. Klüppelberg), pp. 63–107. Boca Raton: Chapman and Hall–CRC.
Lauritzen, S. L. and Spiegelhalter, D. J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Statist. Soc.* 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, Aalborg.
——— (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, **17**, 31–57.
Matúš, F. (1999) Conditional independences among four random variables III: Final conclusion. *Combin. Probab. Comput.*, **8**, 269–276.
Meek, C. (1995) Causal inference and causal explanation with background knowledge. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence* (eds P. Besnard and S. Hanks), pp. 403–410. San Francisco: Morgan Kaufmann.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
Mohamed, W. N., Diamond, I. and Smith, P. W. F. (1998) The determinants of infant mortality in Malaysia: a graphical chain modelling approach. *J. R. Statist. Soc.* A, **161**, 349–366.
Neyman, J. (1923) *On the Application of Probability Theory to Agricultural Experiments: Essay on Principles*. (in Polish) (Engl. transl. D. Dabrowska and T. P. Speed, *Statist. Sci.*, **5** (1990), 465–480).

Ord, K. (1976) An alternative approach to modelling linear systems. Unpublished.
Pearl, J. (1988) *Probabilistic Inference in Intelligent Systems*. San Mateo: Morgan Kaufmann.
———(1993) Graphical models, causality and intervention. *Statist. Sci.*, **8**, 266–269.
———(1995) Causal diagrams for empirical research. *Biometrika*, **82**, 669–710.
———(1998) Graphs, causality, and structural equation models. *Sociol. Meth. Res.*, **27**, 226–284.
———(2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
Preston, C. J. (1973) Generalised Gibbs states and Markov random fields. *Adv. Appl. Probab.*, **5**, 242–261.
Richardson, T. S. (1996) Models of feedback: interpretation and discovery. *PhD Thesis*. Carnegie-Mellon University, Pittsburgh.
———(1998) Chain graphs and symmetric associations. In *Learning in Graphical Models* (ed. M. Jordan), pp. 231–260. Dordrecht: Kluwer.
———(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*. New York: Wiley.
Roberts, G. O. and Tweedie, R. L. (1996) Exponential convergence of Langevin distributions and their discrete approximation. *Bernoulli*, **2**, 341–364.
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. *Math. Modllng*, **7**, 1393–1512.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Speed, T. P. (1979) A note on nearest-neighbour Gibbs and Markov distributions over graphs. *Sankhya* A, **41**, 184–197.
Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993) Bayesian analysis in expert systems (with discussion). *Statist. Sci.*, **8**, 219–283.
Spirtes, P. (1995) Directed cyclic graphical representations of feedback models. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence* (eds P. Besnard and S. Hanks), pp. 491–498. San Francisco: Morgan Kaufmann.
Spirtes, P., Glymour, C. and Scheines, R. (1993) *Causation, Prediction and Search*. New York: Springer.
Spirtes, P., Meek, C. and Richardson, T. S. (1995) Causal inference in the presence of latent variables and selection bias. In *Proc. 11th Conf. Uncertainty in Artificial Intelligence* (eds P. Besnard and S. Hanks), pp. 403–410. San Francisco: Morgan Kaufmann.
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 Pap. 6th Int. Wrkshp AI and Statistics, Jan. 4th–7th, Fort Lauderdale* (eds 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 equation modeling tool. *Sociol. Meth. Res.*, **27**, 182–225.
Spitzer, F. (1971) *Random Fields and Interacting Particle Systems*. Washington DC: Mathematical Association of America.
Strotz, R. H. and Wold, H. O. A. (1960) Recursive versus nonrecursive systems: an attempt at synthesis. *Econometrica*, **28**, 417–427.
Studený, M. and Bouckaert, R. R. (1998) On chain graph models for description of independence structures. *Ann. Statist.*, **26**, 1434–1495.
Verma, T. and Pearl, J. (1990) Equivalence and synthesis of causal models. In *Proc. 6th Conf. Uncertainty in Artificial Intelligence* (eds P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer), pp. 255–270. Amsterdam: North-Holland.
Wermuth, N. (1992) Block-recursive regression equations (with discussion). *Rev. Bras. Probab. Estatist.*, **6**, 1–56.
Wermuth, N., Cox, D. and Pearl, J. (1994) Explanations for multivariate structures derived from univariate recursive regressions. *Technical Report 94-1*. University of Mainz, Mainz.
———(1999) Explanations for multivariate structures derived from univariate recursive regressions. *Technical Report*. University of Mainz, Mainz.
Wermuth, N. and Lauritzen, S. L. (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. R. Statist. Soc.* B, **52**, 21–72.
Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.
Wold, H. O. A. (1953) *Demand Analysis*. New York: Wiley.
———(1954) Causality and econometrics. *Econometrica*, **22**, 162–177.
Wright, S. (1921) Correlation and causation. *J. Agric. Res.*, **20**, 557–585.

Discussion on the paper by Lauritzen and Richardson

A. P. Dawid (*University College London*)

There are three intertwining strands to this paper.

(a) The authors point out, by examples, that the semantics of chain graph models of conditional independence (involving only observable variables) are not the same as the semantics of directed acyclic graph (DAG) models, even after possible marginalization over and conditioning on unobserved variables.
(b) They describe some data-generating processes that lead to chain graph models (although typically only as an asymptotic equilibrium).
(c) They consider ways in which intervention in a system may affect the underlying distribution, and how this might be modelled.

The first point should not really be a surprise—after all, why should the two different representations be equivalent? The fact remains that practitioners who are less thoughtful than the authors (i.e. all of us) can all too easily fall into the error of using a chain graph model when what is needed is a DAG with unobserved variables (or some other graphical representation). The authors have done a valuable service by pointing out the problems and misunderstandings that this mistake can bring about.

There is, however, a crucial omission from this paper: nowhere does it provide a clear statement of how we can query a chain graph model to extract the conditional independence statements that it implies. Without a clear understanding of this procedure, it is difficult to follow the authors through their analyses of the conditional independence properties of their graphs. The missing statement (based on the so-called 'moralization criterion') can be found, for example, in section 5.4 of Cowell *et al.* (1999). For completeness, I give it here.

Let A, B and C be three subsets of the variables V whose joint distribution is represented by a chain graph $\mathcal{G}$. We first restrict attention to the subgraph induced by the smallest ancestral set containing $A \cup B \cup C$, where a set of variables is termed ancestral if, whenever it contains a variable v, it also contains all parents and neighbours of v in $\mathcal{G}$. In that subgraph, we add an edge (if necessary) between two nodes if they have children in a common chain component ('moralization'), and then remove all arrow-heads. Then we can infer $A \perp\!\!\!\perp B | C$ if, in the resulting undirected graph, every path from a node in A to a node in B intersects C. So long as the joint density $f(\cdot)$ of all the observations is everywhere positive, this Markov property is logically equivalent to the existence of a factorization as displayed in equations (2) and (3). In fact, equation (3) can be simplified: since $\mathrm{pa}(\tau)$ is a complete set in $(\mathcal{G}_{\tau \cup \mathrm{pa}(\tau)})^m$, the normalizing constant Z can be absorbed into one of the ϕ-terms and so need not be explicitly included.

The second strand of this paper, describing underlying data-generating processes, is important in three different ways. First, it is essential for many probabilistic and statistical tasks to have a way of simulating from a specified model. Secondly, an understanding of how the model arises as for example the equilibrium distribution of a well-defined process is invaluable as an aid to an interpretation of what the model is actually saying. And, thirdly, such generating processes can be used, as described in Section 6.4 of the paper, to extend the model to situations involving interventions—so interweaving with the third strand. The authors use this approach to develop their formula (18), which can be regarded as a canonical way of constructing an interconnected collection of models (describing the effects of an intervention to set X_A, for various choices of A), using as starting-point a pair $(\mathcal{G}, P)$, where distribution P is Markov with respect to chain graph $\mathcal{G}$. It should be emphasized that both these ingredients are required to define this 'canonical extension'. If P is Markov with respect to $\mathcal{G}_1$, and $\mathcal{G}_2$ is Markov equivalent to $\mathcal{G}_1$ (as described in proposition 1 of the paper), then of course P is Markov with respect to $\mathcal{G}_2$. Nevertheless the associated collection of intervention models, given by equation (18), will differ. Which—if either—of these intervention collections corresponds to the way that the world actually works cannot be a matter of algebraic manipulation, but of empirical investigation. In particular, in interpreting equation (5) we must regard the two sides as defined quite independently of one another, the left-hand side being determined by how the world actually works, and the right-hand side by pushing symbols around. Since in general there is no good reason to expect equality between these two very different things, to say that a DAG $\mathcal{D}$ is causal for P is a very strong requirement, even when P is Markov with respect to $\mathcal{D}$. Likewise, when P is Markov with respect to a chain graph, there is absolutely no reason why formula (18) should describe the actual effects of interventions: it is merely a mathematically convenient suggestion, possibly worth further empirical investigation.

Now we do not have to think in terms of generating processes to make sensible suggestions for modelling intervention. Instead, we might attempt to modify the graph to incorporate such interventions. For DAG models, this approach has been followed by Spirtes *et al.* (1993), Pearl (2000), section 3.2.2, and Lauritzen (2000) and further developed by Dawid (2002a, b). It extends readily to more complex graphical representations such as chain graphs.

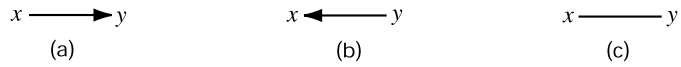

Fig. 7. Three equivalent chain graphs

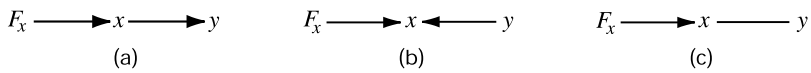

Fig. 8. Corresponding augmented graphs

Thus consider the three simple chain graphs displayed in Fig. 7. These are trivially Markov equivalent, none of them putting any constraint whatsoever on the joint density $f(x, y)$ of x and y. Graphs 7(a) and 7(b) correspond respectively to the always available factorizations $f(x, y) = f(x)\, f(y|x)$, and $f(x, y) = f(y)\, f(x|y)$, whereas graph 7(c) represents the trival factorization $f(x, y) = f(x, xy)$.

To supply a model for the effects of an intervention at x, we introduce a new *intervention node* F_x, together with an arrow from F_x into x. The resulting *augmented graphs* are displayed in Fig. 8.

The possible states of F_x are the same as those of x, together with an additional state $\varnothing$. Conditionally on $F_x = \varnothing$, the joint density $f(x, y|\varnothing)$ is taken to be that corresponding to (x, y) arising naturally. A value $x^* \neq \varnothing$ for F_x is interpreted as corresponding to an intervention to set x to the value x^*. Obviously, given $F_x = x^*$, the distribution of x must be degenerate at x^*. The question is: 'How should we model, in a canonical way, the resulting distribution of y?'.

Even through F_x is not a regular random node, let us apply standard graphical semantics to the augmented graphs. Using proposition 1 of the paper we then see that graphs 8(a) and 8(c) are equivalent—but these are not now equivalent to graph 8(b). Correspondingly, using the moralization criterion we find that for graphs 8(a) and 8(c) the associated graphical model implies $y \perp\!\!\!\perp F_x | x$, whereas for graph 8(b) it implies $y \perp\!\!\!\perp F_x$. The former property implies $f(y|x = x^*, F_x = x^*) = f(y|x = x^*, F_x = \varnothing)$. That is, the density of y when we intervene to set $x = x^*$ is being taken to agree with the *conditional* density $f(y|x^*)$ calculated from the natural joint distribution. However, the property $y \perp\!\!\!\perp F_x$ embodied in graph 8(b) implies $f(y|F_x = x^*) = f(y)$, i.e. the interventional distribution of y is now being taken to agree with its natural *marginal* distribution.

In general neither of these assumptions is obviously preferable to the other, and in applications either or both may fail to provide a good model for how the world actually behaves. The advantage of using augmented graphs such as those in Fig. 8 is that they display with great clarity exactly what is being assumed, so laying this open to reasoned criticism and empirical testing.

Consider now the more complex chain graph model displayed in Fig. 9(a). Again we can introduce a node F_x to model an intervention at x. The resulting augmented graph is displayed in Fig. 9(b). The broken lines represent the moralization edges required whenever either x or y is included in the variables whose conditional independence is being queried.

From Fig. 9(b) we can read off the following properties:

(a) $\perp\!\!\!\perp \{a, b, F_x\}$ (mutual independence of a, b and F_x)—i.e. a and b are independent, and moreover are unaffected by (and do not affect) whether x arises randomly or is set by intervention;
(b) $y \perp\!\!\!\perp (F_x, a) | (b, x)$—the distribution of y given b and x is unaffected by further conditioning on a and moreover does not depend on whether x arises randomly or is set by intervention;
(c) $x \perp\!\!\!\perp b | (a, y, F_x)$—the distribution of x given a and y is unaffected by further conditioning on b, when x arise randomly (the same property under intervention at x holds trivially).

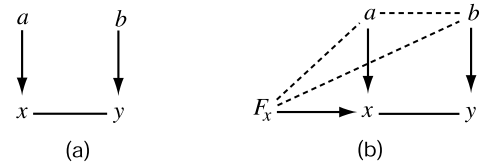

Fig. 9. (a) Chain graph and (b) augmentation

Once again, these properties may or may not be appropriate for any particular physical process that we are trying to model, but they are reasonably natural, and—most importantly—are clearly and explicitly displayed in the augmented graphical representation.

The above construction, extended to allow multiple intervention nodes each pointing into its associated regular node, can clearly be applied to an arbitrary chain graph. It is then not difficult to check that formula (18) follows directly from the moralization criterion applied to the associated augmented graph, without any need to consider underlying generating mechanisms.

The method can be applied still more generally, to other graphical representations such as reciprocal graphs (Koster, 1996), annotated graphs (Paz *et al*., 2000) and Markov chain graphs (Koster, 2000), as well as to alternative semantic interpretations of chain graphs (Andersson *et al*., 1996). Appropriate canonical formulae to describe the effects of interventions in such other graphical frameworks can then be derived from the relevant semantics, applied to the augmented graph.

It will be obvious that I have been greatly stimulated by this paper, and it gives me much pleasure to propose a vote of thanks to its authors.

D. R. Cox (*Nuffield College, Oxford*)

Much as I admire both Professor Lauritzen and Professor Richardson's work, I think that the present paper, while containing interesting discussion, is marred by taking too limited a view of the notion of chain independence structures. Take a simple case of four nodes representing two response variables Y_1 and Y_2 and two explanatory variables X_1 and X_2. Then a missing (directed) edge between, say, Y_1 and Y_2 represents the independence $Y_1 \perp\!\!\!\perp X_2 | Y_2, X_1$, i.e., when we consider the dependence of one component response on explanatory variables, we take the other response variables as explanatory. Now this idea induces an interesting and important class of models, developed by Lauritzen and Wermuth (1989) in their landmark paper. But, as indeed the present authors recognize, this is not the only possibility and Cox and Wermuth (1993, 1996) suggested a second type of graph with directed and undirected edges, which they represented with broken lines. This covers the possibility that the missing edge represents instead $Y_1 \perp\!\!\!\perp X_2 | X_1$, i.e. considering in the conditioning set only truly explanatory variables.

It seems to me that, however we choose to represent them, we need to recognize such structures as having both a simple generating process and an interpretation that are different from but at least equally legitimate as the first. Indeed the overwhelming majority of mainstream statistical discussion is of the second not the first structure. Recall, for instance, the multivariate linear model and the notion of seemingly unrelated regressions. It is true that for the second kind of structure the theory for the linear model does not generalize as elegantly as in the former structure.

Consider the further implications. We may have data that can be represented in five blocks in a sequence of dependences and each block containing in general several variables on an equal footing: X_0, base-line intrinsic variables, X_B a set of early risk factors, X_C and X_C^*, two explanatory variables of immediate concern, X_I, a set of intermediate responses and Y, the ultimate response of interest. We ask 'what is the effect of X_C on response?'. The authors are right to criticize implicitly the statistical literature for failing to address this kind of question, but I believe that workers in many applied fields are clear that the right procedure, in the absence of X_C^*, is to condition on X_0 and X_B and to marginalize over X_I; this would normally be expressed via the inclusion and exclusion of variables as explanatory in some form of regression analysis. An important role of X_I is, however, to explain the pathways via which any effect of X_C is produced, but that is a different issue which is not considered in the following.

But what of X_C^*? Suppose to be specific that X_C and X_C^* represent sodium and potassium levels in the blood and Y is a medical outcome. We consider a hypothetical unit change in sodium level, X_C; what happens to Y? There seem to be at least four possible answers to that question, depending on what happens to X_C^*.

First we might fix X_C^*, i.e. demote X_C^* to being a component of X_B. This appears to be the only approach that the authors consider. It is in the spirit of the representations that they favour. In some contexts, e.g. if the variables represented sodium and potassium intake rather than levels in the blood, it might be preferred. In others it would involve enforcing a lack of equilibrium in some underlying process.

Secondly we might allow X_C^* to vary in accordance with the relationship obtaining in the data under analysis. This would involve promoting X_C^* to be a component of the intermediate variable X_I. This might make sense if the regression of potassium on sodium level were observed in a longitudinal study within individuals but almost certainly not if the regression were based only on a variation between individuals because of the absence of longitudinal data.

A third possibility is that a separate investigation is needed of the changes of X_C^* consequent on enforced changes of X_C. For least squares regression problems, this amounts to using the equation

$$\beta_{YX_{C^\circ}X_C^*} = \beta_{YX_C.X_C^*} + \beta_{YX_{C^*}X_C}\gamma_{X_CX_C^*},$$

where $\beta_{YX_{C^\circ}X_C^*}$ is the regression coefficient of Y on X_C when X_C^* varies in the new way specified and the regression coefficient $\gamma_{X_CX_C^*}$ refers, in general, to the new system different from the initial one. See Cox (1984) and for an application Anderson and Cox (1950).

The fourth interpretation is to declare the initial question meaningless, i.e., at least until the interplay between X_C and X_C^* is better understood, only the dependence of Y on the complex as a whole can be studied.

Note that, if the final response Y has several components, it seems particularly artificial in the present context to include some of these components as, in effect, explanatory variables.

Which of these interpretations is to be preferred is a subject-matter issue and cannot be settled by seeing which fits best into some preordained mathematical structure.

I have much pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Svend Kreiner (*University of Copenhagen*)

There are many interesting and provocatory points in this paper on the incommensurability of directed acyclic graphs (DAGs) and chain graph models (CGMs). It is certainly important to remember that different types of model often are inconsistent and that we should not attempt to interpret features of CGMs in terms of simple parsimonious DAGs. Section 5 thus discards CGMs where the use of CGMs is discussed in terms of misuse and naïvety, but they are revived in Section 6, where it is shown that different types of data-generating feed-back mechanisms in infinite DAGs may result in models that at least may be approximated by CGMs. The arguments in these two sections and the following discussion of statistical analysis by CGMs and DAGs appear to be based on the basic assumption that a data-generating mechanism must always be a DAG: that DAGs are more than a statistical model or at least are closer to reality than other statistical models. The assumption is not stated explicity in this and other sources on causal modelling, but if this is not what the authors mean then the arguments of the paper reduce to the almost empty statement that different models may be inconsistent—an argument that does not justify the judgmental terms used by the authors.

It is unfortunate that we must infer these viewpoints and it would be useful if the authors for once would come clean on this point so that we can start discussing the pros and cons of this point of view. It is also unfortunate that the belief in the fundamental nature of DAGs is not supported with examples from actual research. Instead causal DAGs are usually discussed in loose epidemiological terms of risk factors, diseases and symptoms where a causal interpretation of relationships is relatively easy. A discussion of symmetric relationships must look beyond epidemiology to other research areas. For a no longer recent but still valid discussion of symmetrical relationships in sociology see for instance Rosenberg (1968). In addition to the same types of relationships discussed by Lauritzen and Richardson, Rosenberg also discussed relationships where the existence of a data-generating DAG—with or without feed-back—is not obvious, e.g. functional relationships of elements of units, relationship between parts of a common system and fortuitous relationships rooted in accidental events at macro levels.

Peter Green (*University of Bristol*)

The authors' thesis that their data generation processes provide valid interpretations of chain graphs, supporting intervention, is characteristically original and intriguing; central to this is the relationship between intervention by replacement in a data generation process with ordinary conditioning in the equilibrium distribution, and that is what I want to comment on.

This relationship is discussed in Sections 6.3 and 6.4, and directly extends that for undirected graphs in Section 6.2; all that I want to say about the extension to chain graphs is that surely convergence to equilibrium could occur concurrently rather than sequentially across the chain components.

Section 6.2 provides helpful examples, rather than a full analysis of the connection between intervention and conditioning; the present discussion is only slightly less incomplete. From these examples, we have the impression that reversibility is the key. The first four processes listed—the

systematic and random Gibbs samplers, the time reversible Markov dynamics and the Langevin diffusions—are all reversible (in the case of the first, at least at the level of individual updates). It is easy to contrive examples of non-reversible generating processes where condition (14) fails, i.e. where intervention by replacement does not lead to ordinary conditioning. However, further study shows that reversibility is not enough.

The vector diffusion (12), $X(t + \mathrm{d}t) = X(t) + CX(t)\,\mathrm{d}t + \mathrm{d}Z(t)$ with $\mathrm{var}\{\mathrm{d}Z(t)\} = \Lambda\,\mathrm{d}t$, is useful in exploring things further. When $\Lambda = I$, this is reversible if and only if C is symmetric, precisely the requirement for equation (14) to hold, as identified by proposition 5. But, for general Λ, the process is reversible if and only if $C\Lambda = \Lambda C^{\mathrm{T}}$. As long as Λ is diagonal, the impact is straightforward—proposition 5 can be easily modified, and reversibility continues to imply condition (14). However, if Λ is not diagonal, this breaks down: your cannot simultaneously have condition (14) and obtain the correct equilibrium without intervention. Diagonality of Λ is the same as saying that perturbations to the individual variables are (conditionally) independent.

A similar complication arises if you take a broader view of the Gibbs sampler, allowing block updates and directional sampling: again, condition (14) holds only when the perturbations are independent. When perturbations are independent, intervention by replacement is equally *intervention by conditioning*, so the connection to ordinary conditioning in the equilibrium is perhaps unsurprising.

Curiously, the discrete time version of diffusion (12) behaves differently: suppose that $X_{t+1} \sim N(AX_t, \Lambda)$. Reversibility ($A\Lambda = \Lambda A^{\mathrm{T}}$) and independent perturbations (Λ diagonal) are not enough: for condition (14) you need A diagonal as well, when the system decouples completely. Perhaps the advantage of continuous over discrete time processes is of more than 'intuitive appeal'.

Bill Shipley (*Université de Sherbrooke*)

I present these comments as someone whose interest is in using graphical models to analyse and interpret empirical data. Although statisticians for many years have studied chain graph models, such models are only now beginning to be applied to empirical data. Because most empirical studies are based—implicitly or explicitly—on causal hypotheses, it is important to describe clearly the causal interpretations that are expressed in a chain graph. It is common in empirical research to propose multivariate causal hypotheses in which some of the causal relationships between variables are unknown and so appear symmetric. It is intuitively appealing to represent such undefined associations as undirected edges. As Lauritzen and Richardson show in this paper, intuition can be a poor guide. Different causal interpretations of the same chain graph (whether the undirected edge represents an unmeasured parent, a selection process or a feed-back relationship) imply different patterns of conditional independence. Since each such causal interpretation can be expressed as a directed acyclic graph or a directed cyclic graph—perhaps involving latent nodes—one can ask 'is there any causal process for which a chain graph is preferable to a directed graph?'.

I suspect that the answer is 'no'. However, one aspect of chain graph models that was not discussed in this paper, but which might provide an affirmative answer, is in the testing of the model against empirical data. A directed acyclic graph can be tested quite generally by obtaining the basis set of d-separation relationships implied by the model followed by tests of the implied (conditional) independences in the data. When some d-separation claims involve latents (which can be represented by an undirected edge in a chain graph) then such tests cannot be performed directly since this would require conditioning on an unmeasured variable. If the model is a directed cyclic graph then there might be independences that are not captured by d-separation. The only available tests in such cases involve the methodology of structural equation modelling and these tests require some assumptions that are often difficult to meet in practice. Might it be possible to obtain a basis set of the conditional independences implied by a chain graph relative to a specific causal interpretation of the undirected edges that would permit the more general tests of conditional independence without requiring that we condition on unmeasured variables?

J. T. Kent (*University of Leeds*)

The chain graph models of this paper are built out of two ingredients: undirected graphs and directed acylic graphs. Both of these have important applications in image analysis where the vertices are usually a set of sites or pixels arranged in a rectangular array. In a Markov random field, each site is linked to nearby neighbours in an undirected graph. However, a Markov mesh model is a directed acyclic graph in which each site has parents below and to the left. In both cases these are generally models of

convenience without any scientific justification. Indeed a common modelling strategy is to increase the size of neighbourhood until the fit is satisfactory. One question is whether it is possible to combine Markov random fields and Markov mesh models to obtain more general chain graph models with fruitful imaging applications.

Jim Q. Smith (*University of Warwick, Coventry*)

I congratulate the authors for their thought-provoking and well-written paper. I would like to make two observations, both about causality or manipulation in general.

They demonstrate lucidly that models manipulated naturally using structural equation modelling do not necessarily give the same rules as models naturally manipulated by augmenting the dynamics generating their equilibrium distribution. Most of the Bayesian graphical models that I build encode expert judgments. When eliciting such judgments it is often most natural to ask the expert to consider what would happen if certain covariates were held fixed and others possibly randomized. Thus in models of the spread of nuclear contamination engineers have a good appreciation of how the components of their plant behaved in test conditions, physicists understand how the plume of released contamination should move in a given wind field in stable atmospheric conditions and various species of plant are exposed to radiation in controlled randomized trials to measure the nature and variability of their absorption. Manipulated submodels are thus usually the primitive inputs to these composite networked Bayesian models. The idle composite and other manipulated composites are obtained by combining these components appropriately.

This is the reverse of that studied here. Could the limits of applicability of 'off-the-shelf' families like chain graphs be largely explained by the possibility that our knowledge and reasoning are most easily performed and articulated in manipulated and not idle systems? Manipulated systems express substantive knowledge more often than the idle uncontrolled systems and we should make assertions about what we expect in controlled environments and only then make the bold extrapolation to the idle system. Taking this stance we become less worried that a family of graphical models does not fit many generating processes. Rather our idle model is customized to the science of the application.

Second, I am uncomfortable with the apparent suggestion that graphical models should be first fitted incorporating no other background knowledge (e.g. about the causal order of variables—see the end of Section 8.3) and only then sorted by their contextual plausibility. This approach seems to lift conditional independence hypotheses into an unjustified exulted position. You search through an enormous class of models assuming that various combinations of conditional independence relationships hold, many of which may be contextually unexplainable or implausible, ignoring all other background information you have. You identify the graphical models that fit best and then accept the most likely if you can match it with an explanation. Where do causal hypotheses which apparently require some specific conceptual frame fit into this rather arbitrary and non-specific exploratory data analysis?

Have I misinterpreted the authors? If so, could they perhaps convince me with a contextual example that leads to intellectually coherent reasoning?

The following contributions were received in writing after the meeting.

Jan T. A. Koster (*Erasmus University, Rotterdam*)

I would like to draw attention to a consequence of the proposed causal interpretation of chain graphs which might worry some researchers wishing to extend their use of chain graphs from the ordinary statistical domain to the more elusive causal domain, where intervention conditional densities are used to model the consequences of concrete (e.g. experimental) interventions. As my point can already be made clear for the special case of undirected graphs, I shall just consider this case. Suppose that G is an undirected chordal (i.e. no cycles of length 4 or greater) graph, and let $\mathcal{D}$ be the (non-empty) class of directed acyclic graphs (DAGs) that are Markov equivalent to G. Frydenberg (1990) has shown that any such DAG has the same adjacencies as G, and no immoralities. Many researchers would accept G as a safe and succinct representation of the class $\mathcal{D}$ whenever it is believed that some $D \in \mathcal{D}$ is causal for P—a view which the authors would probably call 'epistemological'. Suppose that f is the density of P, where P is G Markov, and let $D \in \mathcal{D}$. Since P is D Markov as well, f can be factored according to formula (1); hence the conditional density of $X_{V\setminus A}$ given X_A can be expressed as

$$f(x_{V\setminus A}|x_A) = \prod_{v\in V\setminus A} f(x_v|x_{\mathrm{pa}_D(v)}) \prod_{v\in A} f(x_v|x_{\mathrm{pa}_D(v)}) f(x_A)^{-1}.$$

Assuming that D is causal for P, the intervention conditional density induced by D is denoted by $f(x_{V\setminus A}||x_A)$ and is given by formula (5) of the paper. So $f(x_{V\setminus A}||x_A) = f(x_{V\setminus A}|x_A)$ holds, if and only if $f(x_A) = \Pi_{v\in A} f(x_v|x_{\mathrm{pa}_D(v)})$, if and only if, for all $v \in A, \mathrm{pa}_D(v) \subseteq A$ and P_A is D_A Markov, if and only if $\mathrm{pa}_D(A) = \emptyset$ (since this entails that P_A is D_A Markov). Since the intervention conditional density for the undirected graph G satisfies $f(x_{V\setminus A}||x_A) = f(x_{V\setminus A}|x_A)$ by definition, under the epistemological interpretation of G this will only be correct as a model for a concrete intervention $X_A \leftarrow x_A$, if for the true causal DAG $D \in \mathcal{D}$ it holds that $\mathrm{pa}_D(A) = \emptyset$. If $\mathrm{ne}_G(A) \neq \emptyset$, it follows from corollary 4.1 of Andersson *et al.* (1997) that, for some $D \in \mathcal{D}$, $\mathrm{pa}_D(A) \neq \emptyset$. Also, unless $\mathcal{D} = \{G\}$ (i.e. G has no edges), for some intervention set $A \subseteq V$ it will be the case that $\mathrm{pa}_D(A) \neq \emptyset$, where $D \in \mathcal{D}$ is the true causal DAG. Of course, the issue does not arise if G is non-chordal, as in that case $\mathcal{D} = \emptyset$, i.e. there are no DAGs which are Markov equivalent to G.

David Madigan (*Rutgers University*), **Steen A. Andersson** (*Indiana University, Bloomington*) **and Michael D. Perlman** (*University of Washington, Seattle*)

We congratulate the authors for this insightful and informative paper. Particularly instructive is their demonstration of possible misinterpretations of chain graph (CG) Markov models, which here and elsewhere we refer to as LWF (Lauritzen, Wermuth and Frydenberg) CG models.

In Section 8.4 the authors note that an alternative Markov property (AMP) has been developed for CGs by Andersson *et al.* (1996, 2001) (in fact a special case of the Cox and Wermuth (1993, 1996) family of joint response CG models). The authors point out that, although the AMP associated with graph CG_3 in Fig. 1(b) coincides with that obtained by marginalization over the hidden variable h in graph DAG_4 in Fig. 2(a), this does not generally hold for CGs under the AMP (or LWF) interpretation. None-the-less, a direct, in fact linear, representation is possible for *Gaussian* CG models under the AMP interpretation; this representation can be interpreted as a direct data-generating process provided that correlated errors are permitted.

For graph CG_3 this representation takes the form (cf. Anderson *et al.* (2001), section 1)

$$\left.\begin{aligned} X_a &= \varepsilon_a, \\ X_b &= \varepsilon_b, \\ X_c &= \beta_{ca} X_a + \varepsilon_c, \\ X_d &= \beta_{db} X_b + \varepsilon_d, \end{aligned}\right\} \tag{24}$$

where β_{ca} and β_{db} are non-random scalars and where ε_a, ε_b and $(\varepsilon_c, \varepsilon_d)$ are mutually independent random errors with zero means, ε_a and ε_b have univariate normal distributions with arbitrary variances and $(\varepsilon_c, \varepsilon_d)$ has a bivariate normal distribution with *unspecified* covariance matrix. When $\mathrm{corr}(\varepsilon_c, \varepsilon_d) \neq 0$, (X_a, X_b, X_c, X_d) satisfies the AMP conditions for CG_3 but not the LWF conditions.

Such a linear representation remains valid for a general CG G: the AMP for G is equivalent to the set of conditional independences satisfied by a block-recursive Gaussian linear system naturally associated with G. Each variate is expressed as a linear function of its parents in the CG, together with a Gaussian error term. The errors are independent across blocks, whereas within each block the errors are (possibly) correlated according to the undirected subgraph in the block, i.e. according to a Gaussian *covariance selection model* (Dempster (1972) and Lauritzen (1996), section 5.2)—see Andersson *et al.* (2001), remark 5.1, for details. Therefore a general AMP CG model with Gaussian errors can be interpreted as arising from a (linear) data-generating process exactly to the extent that Gaussian covariance selection models admit such an interpretation (or, if covariance selection models simply are viewed as primitives). Thus, if each undirected block subgraph is decomposable and hence Markov equivalent to some acyclic directed graph (Andersson *et al.*, 1997), the within-block errors can be expressed as linear combinations of uncorrelated errors; hence the entire block-recursive Gaussian linear system can be expressed in terms of uncorrelated errors. For example, in expressions (24) $(\varepsilon_c, \varepsilon_d) \sim (\delta_c, \gamma\delta_c + \delta_d)$ (or vice versa), where $\mathrm{corr}(\delta_c, \delta_d) = 0$.

Michael D. Perlman (*University of Washington, Seattle*), **Steen A. Andersson** (*Indiana University, Bloo-mington*) **and David Madigan** (*Rutgers University*)

It is of interest to note the difference between the Lauritzen, Wermuth and Frydenberg (LWF) Markov property and alternative Markov property (AMP) for the chain graph CG* obtained from CG_3 by adding an undirected edge between nodes a and b. Under the LWF interpretation CG* entails the two conditional independences $a \perp\!\!\!\perp d|\{b, c\}$ and $b \perp\!\!\!\perp c|\{a, d\}$, which remain unchanged if (a, b) and (c, d) are

interchanged. Thus CG* is Markov equivalent to the fully undirected graph obtained by converting the two arrows $a \rightarrow c$ and $b \rightarrow d$ into undirected edges, as well as to the CG obtained by reversing the two arrows, thereby losing the intuitive notion of causality conveyed by these two arrows in CG*. By contrast, the AMP for CG* entails the two conditional independences $a \perp\!\!\!\perp d|b$ and $b \perp\!\!\!\perp c|a$, which do not allow the interchange of (a, b) and (c, d). Of course, no graphical Markov property can endow a simple bivariate graph $a \rightarrow b$ with a causal interpretation.

In Section 5.3 it is shown that the most parsimonious LWF CG model that is compatible with the blocking shown in Fig. 4(b) and containing the Markov model determined by graph DAG_1 in Fig. 4(a) is the saturated model given by graph CG_1. However, a more parsimonious model that is compatible with the blocking and containing the DAG_1 model is that given by the CG $a \rightarrow x—y$ under the AMP interpretation, since this AMP CG is in fact Markov equivalent to DAG_1. Of course, given a general DAG model and a compatible blocking, there need not be a Markov equivalent AMP or LWF CG model with the specified blocking. An example is the DAG obtained from CG_3 by converting $c—d$ into $c \rightarrow d$, together with the blocking of CG_3.

The rationale for *substantive* ordered blocking of variates is carefully examined by the authors in Section 5.3. A second rationale for blocking is the occurrence of *symmetries* among variates; these symmetries induce blocking in a natural way. For example, Andersson and Madsen (1998) and Madsen (2000) treated DAG models whose vertices are themselves multivariate blocks, with symmetries present within these blocks.

Christian P. Robert and Jean-Michel Marin (*Université Paris Dauphine*)

As naïve newcomers to the field, we fail to perceive the incentive of using partially directed graphs for statistical or inferential purposes. In particular, the absence of real examples in the paper is a hindrance to understanding what is known, unknown or sought: are the joint, marginal or conditional distributions known? Are the node realizations observed? For instance, taking graph (b) of Fig. 1, simple conditioning leads us to the directed representation

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

Obviously, if $f(d|a, b)$ is *unknown*, we understand (and appreciate) the Gibbs data-generating process using only the conditionals $f(c|a, d)$ and $f(d|c, b)$.

If the conditional distributions of the nodes are known, undirected edges do not seem relevant for the statistical processing (e.g. estimation) of the parameters underlying the distributions of the various bits of the graph. For instance, in graph (a) of Fig. 2, if it is known that there is a latent variable h, this information can be incorporated within the model for later processing, as in the EM and Gibbs algorithm. Similarly, if there is a known feed-back structure as in graph (c) of Fig. 2, or the corresponding example of Whittaker (1990), this piece of information should be incorporated during the statistical processing.

However, if the purpose of the study is to determine causal relations (e.g. model choice), is there enough information within the data *per se* to reach a conclusion? Several causal structures can lead to the same likelihood, and thus should lead to the same inference or should not be discriminated. As mentioned in the paper, the representation

$$X = \Gamma X + \Delta\varepsilon$$

is equivalent, likelihoodwise, to the representation

$$X = (I - \Gamma)^{-1}\Delta\varepsilon$$

for the estimation of the parameters Γ and Δ.

At a more technical level, we think that there is a potential danger in using feed-back models in that conditional distributions are not always compatible with a global joint distribution, as discussed in Hobert and Casella (1996).

Paul R. Rosenbaum (*University of Pennsylvania, Philadelphia*)

Graphs are fun, conditional independence is fundamental and Lauritzen and Richardson deserve congratulations for new insights into their interconnections.

Do graphical models clarify causality? Is it useful to imagine causal intervention as changing a line of code in a computer program? In clarifying cause and effect, I am pessimistic about multivariate models, but optimistic about better data. Better data mean better research design and implementation:

(a) articulation of the empirical consequences of scientific theories;
(b) controlled experimentation or the identification of natural laboratories where disturbances are controlled and treatments inflicted haphazardly;
(c) precise collection of detail, including temporal order

(Fisher, 1935; Meyer, 1995; Piantadosi, 1997; Rosenbaum, 1999, 2002; Angrist and Krueger, 2000; Cox and Reed, 2000; Rosenzweig and Wolpin, 2000; Shadish *et al*., 2002).

The canonical example of simultaneous equations is from price theory, with supply and demand determining and being determined by price. The atoms of price theory are single purchases, now available from check-out scanners with longitudinal data for individual customers. With these better data, price and demand changes are temporally ordered, not simultaneous, and real consumers in real stores behave differently from the prediction of supply-and-demand equations (Fader and Hardie, 1996).

Two fields that used structural equations are macroeconomics and statistical genetics, but both have acquired vastly better data, to which they have applied new, intriguing, simple, convincing research strategies—convincing because of simplicity.

Parents deliver to offspring a distribution of genotypes that is exchangeable between their biological siblings, i.e. exchangeable within sibships. Exploiting this fact, together with direct measurement of genotype markers and disease outcomes, Spielman and Ewens (1998) developed a permutation test—a variant of Mantel's (1963) test (Laird *et al*., 1998)—to identify markers closely linked to genes that contribute to disease susceptibility. See also Risch and Merikangas (1996) and Horvath and Laird (1998). In contrast, path analyses typically use neither genotypes nor sibships, confounding genes and environment.

Studying the microfoundations of macroeconomics, Gross and Souleles (2002) used detailed longitudinal data about 230000 credit card accounts, including credit quality scores, to examine Friedman's (1968) permanent income hypothesis which asserts that consumers spend in accord with their expected long-term income. Credit card companies expand individual credit limits by using credit scores and idiosyncratic timing rules. Controlling for credit scores and exploiting idiosyncrasies in timing rules, Gross and Souleles (2002) find that increases in credit limits meaningfully predict immediate increases in consumption, counter to the permanent income hypothesis.

Causes operate through mechanisms—genes, purchases, credit expansions—which take time: treatment precedes effect. With better data, you would see more of it.

Alberto Roverato (*University of Modena and Reggio Emilia, Modena*) **and Guido Consonni** (*University of Pavia*)

This paper raises several interesting and thought-provoking issues. We would like in particular to dwell on the use of background information on the ordering of the variables involved, especially for Bayesian inference. We shall consider the case wherein the data-generating process is a directed acyclic graph (DAG) model and no latent variables are present.

Assume first that the ordering of the variables is known. To perform Bayesian model determination, prior elicitation on the parameters of each model is required. Ordering variables can vastly simplify this process through the imposition of properties such as global parameter independence and prior modularity. We remark that these properties may also be applied with respect to *all* orderings of the variables, although this severely restricts the choice of prior families (essentially Wishart and Dirichlet respectively for Gaussian and multinomial sampling schemes).

We now turn to the more realistic situation wherein the ordering of variables is not assumed to be known. It is well established that model determination procedures confined to undirected graphical models (i.e. chain graph models with only one chain component) are not suitable because there are DAGs that are not Markov equivalent to any undirected graph: indeed the appropriate structure to be considered here is the *essential graph*. Now suppose that we are willing to assume only a partial ordering of the variables, so that only *blocks* of variables can be ordered. If, as discussed by the authors, this background information translates into a chain graph with more than one component, then conceptual difficulties arise, as clearly illustrated in Section 5.3. However, this is only to be expected, since a chain graph is merely replicating to several blocks the (inappropriate) undirected case discussed above. We conjecture that a more suitable approach is to require that edges between vertices belonging to distinct blocks be directed, whereas within blocks both directed and undirected edges should be allowed. In this way both graph DAG_1 and graph CG_1 of Fig. 4 would be simultaneously taken into consideration.

A chain graph is sometimes referred to as a 'DAG of chain components'. Our current research is exploiting the idea of using 'DAGs of essential graphs' to perform Bayesian model determination when only partial prior information on the ordering of variables is available. Our point here is that background information on the ordering of variables may be a powerful tool to structure prior beliefs, while in no way unduly restricting the model space.

Milan Studený (*Academy of Sciences of the Czech Republic, Prague*)

My comment concerns the concept of *data-generating processes* (DGPs). I have difficulty in understanding what is meant by this phrase in general. Nobody has given me a mathematical definition on the basis of which I can decide what is supposed to be a DGP for a particular (graphical) model of conditional independence structure and what is not. In my view, a DGP is simply a mental construct by means of which a statistician explains why a particular model occurs in practice. Thus, it is made up to justify a particular class of models and the criterion of what is a DGP is highly subjective! Cox and Wermuth (1996) wrote about implications of DGPs on an intuitive level whereas Lauritzen and Richardson give more detailed examples of DGPs. I appreciate it very much.

Whereas the concept of a DGP for directed acyclic graphs is clear, the case of undirected graphs (UGs) still depends on intuition. Being a mathematician accustomed to a detailed specification of steps I have difficulty in understanding the instruction '*repeat until equilibrium*' in the 'computer program' in Section 6.1. In my view, this phrase represents an even more complex decision process and to explain it fairly many more details are needed. This is my impression resulting from discussion with Thomas Richardson who tried to give me further details.

Thus, I ask myself, if the aim of the concept of a DGP is to justify the use of a model why not give a simpler explanation of UG models in terms of factorization of a distribution after the cliques? Perhaps this is not an explanation that statisticians are accustomed to, but it is very clear from the point of view of probabilistic expert systems and computer scientists may accept it without difficulty. Indeed, cliques of a UG may be interpreted as *maximal sets of interdependent variables* and this interpretation could be used by a knowledge engineer constructing a UG model and trying to 'extract' structural information from an expert.

Following this line, classical chain graph models could be viewed as natural generalizations of UG models: when pieces of structural information from different experts are combined. As illustrated by example 3.1 in Studený (1998) a basic factorization formula from Section 3.1 is obtained when we have a group of experts, who are 'chronologically' ordered to avoid possible discrepancies between their testimonies, and each of them is asked to give structural information about his or her area of competence and indicate preceding relevant factors. This comment is too short to explain the details but a sketch is given in Studený (1998).

The **authors** replied later, in writing, as follows.

First we would like to thank all the discussants for their interesting comments and suggestions which have given us the opportunity to consider the material in our paper from new angles. Below we give brief replies to the issues that are raised although many would deserve longer and elaborate answers.

Several of the discussants seem to imply that our paper is meant to advocate for or against the application of chain graphs and a specific interpretation of these. However, our primary intent in writing this paper was simply to clarify which explanations are and are not compatible with certain chain graph models.

Although we do see directed acyclic graphs (DAGs) as underlying most causal explanations, when described in sufficient detail, such a description may be useless for practical purposes, particularly if there are deterministic or functional relationships between variables. The diffusion process in Section 6.1 may only be viewed as a DAG at an infinitesimal timescale. Dr Kreiner and Dr Studený appear to argue that chain graphs are themselves primitive explanations and do not require further elaboration. We do not find this viewpoint convincing: in our view more complex models need more 'explanations', 'mechanisms', 'theories' or 'empirical validations' to justify them. The custom of directing edges in accordance with a prior ordering necessitates such a justification: without an interpretation there is no reason *a priori* why these edges should not be directed in the opposite direction!

Consideration of such explanations is also important when considering which class of graphical models to apply in a given situation. Although we agree with Professor Dawid that the use of policy variables F_x provides a simple means of computing the effects of interventions, it is still important to

consider underlying generating mechanisms. For example, as emphasized by Dr Koster, the theory of intervention for chain graphs as described in Section 6 only makes sense when undirected edges are interpreted in terms of feed-back. We agree with Professor Cox when he argues that there are many different types of intervention that may be considered. However, we think that the intervention formula for chain graphs (equation (18)) accommodates his second suggestion, so it is not correct that the first interpretation is the only we consider.

We are grateful to Professor Green for pointing out that our proposition 5 fails when Λ is not diagonal. Indeed, proposition 5 can be strengthened to say that, if variables are scaled so that Λ has all diagonal elements equal to 1, then equation (14) implies $\Lambda = I$ and $C = -K/2$ when 'intervention by replacement' is made with Z_B having the conditional covariance given $Z_A = 0$. The proof is essentially identical with that given in the paper.

We agree with Professor Cox when he argues that other classes of graph are required in the context of DAG models with hidden variables; see point (c) in Section 5. However, as the examples in Section 5.3 demonstrate, the process of blocking appears to serve no purpose if unmeasured confounders may be present. Consequently, we conclude that blocking only makes sense in contexts where it may be argued that—to a reasonable scientific approximation and with caution—it is either known that there are no unmeasured confounders or it is known precisely which variables are children of a common confounder, so that such variables may be blocked together, possibly thereby ignoring information on temporal ordering.

The proposal of Professor Roverato and Professor Consonni to construct a DAG of essential graphs faces the problem that the directed edges *between* blocks will constrain which essential graphs may occur *within* blocks. For example, in Fig. 4(c), there may be an undirected edge between x and y if and only if a is either a parent of both x and y or neither.

The generating process for chain graphs under the alternative Markov property (AMP), described by Professor Andersson, Professor Madigan and Professor Perlman, helps to clarify the distinction between the two chain graph Markov properties. The generating process for chain graphs under the original Markov property involves feed-back among the observed variables themselves (Section 6). The Gaussian generating process, described by Andersson and his colleagues, may be regarded as presupposing feed-back relationships among error terms. This AMP interpretation crucially requires each observed variable to be a *deterministic* function of its parents in the chain graph together with its error variable. For instance, marginalization over $\{x, y, z\}$ in chain graph CG_4 in Fig. 10(a) does not induce an independence model on $\{a, b, c, d\}$ corresponding to chain graph CG_5 in Fig. 10(b) under the alternative property. (Under the AMP $b \perp\!\!\!\perp d \mid \{a, c\}$ in CG_5 whereas this does not follow in CG_4 from either Markov property.)

The first point of Professor Perlman and his colleagues appears to mix together several concepts: specifically, there appears to be an implicit claim that if a graph is asymmetric then the joint distribution should reflect this asymmetry. It is difficult to see *a priori* why this should be so: for example, Markov equivalent DAGs, if not identical, will always contain edges pointing in different directions, yet represent the same set of distributions. The intervention formulae (equations (5) and (18)) make clear that the asymmetry in the directed edges in the graph is reflected in the *causal intervention* distributions, rather than in the structure of a *passively observed* joint distribution obeying the Markov property for the graph. When Professor Perlman and his colleagues state that no graphical Markov property can endow a bivariate graph with a 'causal interpretation', this is true *if* it is interpreted to mean that a Markov property cannot associate an asymmetric joint distribution with such a graph. However, a *causal* Markov property, as described by an intervention formula, does just this: it associates asymmetric intervention distributions with a directed edge!

To answer Professor Kent's question, it would certainly be possible to construct a true hybrid of Markov mesh and undirected Markov models, by constructing a sequence of chain components corresponding to horizontal lines, with vertical arrows between lines. We cannot speculate on whether

Fig. 10. (a) Chain graph CG_4 and (b) a chain graph CG_5 that is *not* Markov equivalent (under either chain graph Markov property) to the marginal distribution of CG_4 over $\{a,b,c,d\}$

this is a fruitful approach to follow. Chain graphs also form obvious candidates for space–time processes, where chain components are undirected lattices representing space, whereas there are arrows between chain components representing dynamics; see Besag (2002) for recent related work.

To address the concerns raised by Professor Smith, a distinction needs to be made between modelling an idle system about which you have detailed causal knowledge and modelling an idle system when you have information only on time order. The former appears to correspond to the situation that he has in mind, whereas the latter context describes many applications of chain graphs, which are somewhat exploratory in nature. These contexts are quite different: '*a precedes b*' does not imply that '*a causes b*'! The remark at the end of Section 8.3 that Professor Smith finds 'uncomfortable' is merely a restatement of the fact, demonstrated in Section 5.3, that knowledge of time order alone may place no restrictions whatsoever on the class of possible independence models in the presence of confounding or selection mechanisms. Also note that to construct a model for a large system in an idle situation from experimental submodels as described by Professor Smith it is important that there are no unmeasured or unaccounted for confounders in the larger system. Finally, although it is true that model searches are often performed in practice to find the best fitting graphical model, nowhere in the paper do we advocate 'accepting' the output of such a search.

Dr Marin and Professor Robert seem to address issues of computation and inference and query what is known and unknown. Our study does not address these issues but is only concerned with the potential validity of various types of causal interpretations of chain graph models. We are grateful for the reference to Hobert and Casella (1996) which is clearly concerned with issues related to those discussed in Appendix A.

We thank Professor Shipley for his comments on the use of graphical models in empirical data analysis. Regarding his question whether there are causal processes for which chain graphs are preferable to directed graphs, we believe that the type of feed-back system shown in Fig. 6 represents such a process. Any one of the local Markov properties for chain graphs described in Lauritzen (1996) will form a 'basis' of independence statements implying the global Markov property (assuming positivity).

We are happy to agree with many of the points made by Professor Rosenbaum: of course better data are better; longitudinal data are also better than aggregate 'equilibrium data'. However, we make the following points.

(a) Simply because a simple model or research strategy may appear more convincing than a complex one does not make it any more or less true. It is worth bearing in mind Savage's comment that a model should be 'as big as an elephant'; see Lindley (1983). Even with a gigantic 'market basket' data set, presumably the usual issues of selection and unobserved variables, e.g. prices in other stores and price reductions coinciding with advertising campaigns, and must still be addressed.
(b) Drawing causal conclusions from observational data without (implicitly) positing a multivariate model is clearly not possible; for example, multiple regression requires an intricate battery of substantive assumptions in order for a given coefficient to be interpretable causally.
(c) Although it may be simplistic to model a causal intervention in terms of modifying a computer program, it does have the advantage of being unambiguous and precise. This has not always been true of discussions of causality couched in the more traditional terms of 'simultaneity', 'exogeneity' and 'endogeneity'.

Finally, economists to whom we have spoken have said that there were many situations in which it is not possible to find 'microdata', and thus it is not surprising that there continues to be a large amount of empirical research which uses macroeconometric models.

References in the discussion

Anderson, S. L. and Cox, D. R. (1950) The relation between the strength and diameter of wool fibres. *J. Text. Inst.*, **41**, T481–T491.

Andersson, S. A., Madigan, D. and Perlman, M. D. (1996) An alternative Markov property for chain graphs. In *Proc. 12th Conf. Uncertainty in Artificial Intelligence* (eds F. Jensen and E. Horvitz), pp. 40–48. San Francisco: Morgan Kaufmann.

———(1997) A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, **25**, 505–541.

———(2001) Alternative Markov properties for chain graphs. *Scand. J. Statist.*, **28**, 33–85.

Andersson, S. A. and Madsen, J. (1998) Symmetry and lattice conditional independence in a multivariate normal distribution. *Ann. Statist.*, **26**, 525–572.

Angrist, J. D. and Krueger, A. B. (2000) Empirical strategies in labor economics. In *Handbook of Labor Economics*, vol. III, ch. 23. New York: Elsevier.
Besag, J. (2002) Likelihood analysis of binary data in space and time. In *Highly Structured Stochastic Systems* (eds P. J. Green, N. L. Hjort and S. Richardson). Oxford: Oxford University Press. To be published.
Cowell, R. G., Dawid, A. P., Lauritzen, S. and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.
Cox, D. R. (1984) Design of experiments and regression (with discussion). *J. R. Statist. Soc.* A, **147**, 306–315.
Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments*. New York: CRC Press.
Cox, D. R. and Wermuth, N. (1993) Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.*, **8**, 204–218, 247–277.
———(1996) *Multivariate Dependencies: Models, Analysis, and Interpretation*. London: Chapman and Hall.
Dawid, A. P. (2002a) Influence diagrams for causal modelling and inference. *Int. Statist. Rev.*, to be published.
———(2002b) Causal inference using influence diagrams: the problems of partial compliance (with discussion). In *Highly Structured Stochastic Systems* (eds A. Frigessi and S. Richardson). Oxford: Oxford University Press. To be published.
Dempster, A. P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.
Fader, P. S. and Hardie, B. G. S. (1996) Modeling consumer choice among SKU's. *J. Marktng Res.*, **33**, 442–452.
Fisher, R. A. (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.
Friedman, M. (1968) Theory and analysis of consumption and savings functions. In *Economic Statistics and Econometrics* (ed. A. Zellner), pp. 236–253. Boston: Little, Brown.
Frydenberg, F. M. (1990) The chain graph Markov property. *Scand. J. Statist.*, **17**, 333–353.
Gross, D. B. and Souleles, N. S. (2002) Do liquidity constraints and interest rates matter for consumer behavior?: evidence from credit card data. *Q. J. Econ.*, **117**, 149–186.
Hobert, J. P. and Casella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear models. *J. Am. Statist. Ass.*, **91**, 1461–1473.
Horvath, S. and Laird, N. M. (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am. J. Hum. Genet.*, **63**, 1886–1897.
Koster, J. T. A. (1996) Markov properties of non-recursive causal models. *Ann. Statist.*, **24**, 2148–2177.
———(2000) Marginalizing and conditioning in graphical models. *Technical Report EUR/FSW-Soc/2000.02*. Erasmus University, Rotterdam.
Laird, N. M., Blacker, D. and Wilcox, M. (1998) The sib transmission/disequilibrium test is a Mantel-Haenszel test. *Am. J. Hum. Genet.*, **63**, 1915.
Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon.
———(2000) Causal inference from graphical models. In *Complex Stochastic Systems* (eds O. E. Barndorff-Nielsen, D. R. Cox and C. Klüppelberg), ch. 2, pp. 63–107. Boca Raton: CRC Press.
Lauritzen, S. L. and Wermuth, N. (1989) Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.*, **17**, 31–57.
Lindley, D. V. (1983) Parametric empirical Bayes inference: comment. *J. Am. Statist. Ass.*, **78**, 61–62.
Madsen, J. (2000) Invariant normal models with recursive graphical Markov structure. *Ann. Statist.*, **28**, 1150–1178.
Mantel, N. (1963) Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure. *J. Am. Statist. Ass.*, **58**, 690–700.
Meyer, B. D. (1995) Natural and quasi-experiments in economics. *J. Bus. Econ. Statist.*, **13**, 151–161.
Paz, A., Geva, R. Y. and Studený, M. (2000) Representation of irrelevance relations by annotated graphs. *Fund. Inform.*, **42**, 149–199.
Pearl, J. (2000) *Causality*. Cambridge, Cambridge University Press.
Piantadosi, S. (1997) *Clinical Trials: a Methodologic Perspective*. New York: Wiley.
Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
Rosenbaum, P. R. (1999) Choice as alternative to control in observational studies (with discussion). *Statist. Sci.*, **14**, 259–304.
———(2002) *Observational Studies*, 2nd edn. New York: Springer.
Rosenberg, M. (1968) *The Logic of Survey Research*. New York. Basic Books.
Rosenzweig, M. R. and Wolpin, K. I. (2000) Natural "natural experiments" in economics. *J. Econ. Lit.*, **38**, 827–874.
Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
Spielman, R. S. and Ewens, W. J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.*, **62**, 450–458.
Spirtes, P., Glymour, C. and Scheines, R. (1993) *Causation, Prediction and Search*. New York: Springer.
Studený, M. (1998) Bayesian networks from the point of view of chain graphs. In *Proc. 14th Conf. Uncertainty in Artificial Intelligence* (eds G. F. Cooper and S. Moral), pp. 496–503. San Francisco: Morgan Kaufmann.
Whittaker, J. (1990) *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.