# Chain Graph Models and their Causal Interpretations*

Steffen L. Lauritzen† Thomas S. Richardson
Aalborg University University of Washington

August 9, 2001

**Abstract**

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

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

## 1 Introduction

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

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

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

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 1974b; Speed 1979; Darroch *et al.* 1980).

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

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

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

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

## 2 Basic Graphical Concepts and Notation

In the present 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 the present paper.

A *partially directed cycle* in a graph $\mathcal{G}$ is a sequence of $n$ distinct vertices $v_1, \ldots, v_n$, $(n \geq 3)$, and $v_{n+1} \equiv v_1$, such that (a) $\forall i$ $(1 \leq i \leq n)$ either $v_i$ — $v_{i+1}$ or $v_i \leftarrow v_{i+1}$, and (b) $\exists j$ $(1 \leq j \leq n)$ such that $v_j \leftarrow v_{j+1}$.

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

A *chain graph,* (CG), is a graph in which there are no partially directed cycles. A chain graph in which there are no undirected edges is a *directed acyclic graph,* most often abbreviated as DAG.

The *chain components* $\mathcal{T}$ of a chain graph are the 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

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

Minimal complexes play a fundamental rôle 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 article is concerned with the chain graph Markov property defined in Lauritzen and Wermuth (1984, 1989) and Frydenberg (1990). There have been a number of alternative suggestions for associating a Markov property with a chain graph (Cox and Wermuth 1993; Andersson *et al.* 1996), which generally are not equivalent to the above and which are not discussed in detail in the present paper.

Below we give the factorization versions of the Markov properties for DAGs and for chain graphs. For further details, the reader is 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

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

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

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

where each factor further factorizes according to the graph as

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

Here $\mathcal{A}(\tau)$ are the complete sets in $\mathcal{G}_{\tau \cup \mathrm{pa}(\tau)}$ and the factor $Z$ is a normalizer

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

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

## 3.2 The global Markov property and Markov equivalence

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

For subsets of variables $A$, $B$, and $S$, the expression $A \perp\!\!\!\perp B \mid S$ 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 $A \not\perp\!\!\!\perp B \mid S$ 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:

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

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

# 4 Causal Interpretation of DAG Models

This section gives a brief description of the now rather standard causal interpretations associated with a DAG given by Spirtes *et al.* (1993) and Pearl (1993, 1995), largely following Lauritzen (2001). The interpretations are both concerned with their *data generating processes* and associated calculation of *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 modify a given probability distribution. Conditional densities are usually calculated as

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

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

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

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

Other authors have used expressions such as, for example, $P(Y_x = y)$, $P_{\mathrm{man(x)}}(y)$, $\mathrm{set}(X = x)$, $X = \hat{x}$, 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 disctinction is very clearly made, see for example Box (1966) or Cox (1984).

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

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

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

## 4.2 Data generating process for DAG models

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

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

where the assignments have to be carried out sequentially in a well-ordering of the directed acyclic graph $\mathcal{D}$, or partly in parallel, so that at all times, when $X_v$ is about to be assigned a value, all variables in $\mathrm{pa}(v)$ have already been assigned a value. The variables $U_v, v \in V$ are assumed to be independent. For any given probability distribution, there is a multitude of choices for $g_v$ and $U_v$ in the generating process. Deriving results based on 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 as invented in the context of genetics (Wright 1921), and exploited in economics (Haavelmo 1943; Wold 1954) and social sciences (Goldberger 1972). Structural equation models 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 will 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 (4) so that $V = 1, \ldots, p$ and writing

```
For i = 1,...,p;

    ε ← runif;
    x_i ← h_i(x_pa(i), ε);

Return x;
```

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

It is an important aspect of structural equation models that they also specify the way in which intervention is to be 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 (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 DAGs

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

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

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

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

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

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

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

Thus in the case of a DAG there is full harmony between the causal interpretations determined by data generating processes, intervention by replacement, and the causal Markov property associated with the DAG. 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 e.g. Edwards and Kreiner (1983), the hypotheses of interest were in some sense causal, studying relationships beween explanatory and response variables. It is clearly unnatural to try to represent a system of such relations, which are asymmetric, by an undirected graph in which all relations are symmetric.

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

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

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

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

- an unmeasured confounding variable;
- some artefact of the way the sample was selected;

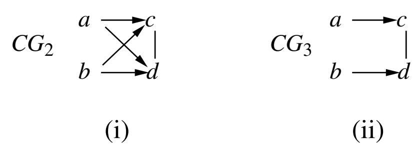



Figure 1: Two examples of chain graphs in which $c$ and $d$ are joint responses to $a$ and $b$.

– a feedback relationship.

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

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

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

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

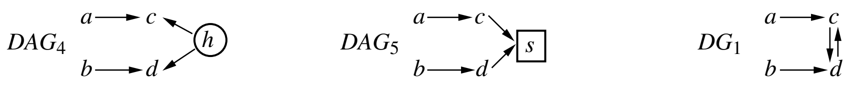
$$f(a, b, c, d) = f(c, d \mid a, b) f(a) f(b).$$

In this sense the model treats $c, d$ as being on an equal footing, as it places no restriction on the form of the conditional density $f(c, d \mid a, b)$. However, when submodels are considered, special attention is required. A submodel such as $CG_3$ in Figure 1 (ii), restricts $f(c, d \mid a, b)$. Under the chain graph Markov property, $CG_3$ implies

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

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

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




(i) (ii) (iii)

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

unmeasured common parent; in the marginal distribution over the remaining variables

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

corresponding to an independence structure different from that of $CG_3$.

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

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

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

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

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

which again does not correspond to $CG_3$.

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

## 5.1 Non-causal associations due to latent variables

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

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

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

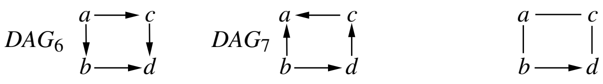
**Proposition 3** *A chain graph $\mathcal{K}$ over the variables $V$ represents the same set of conditional independence relations as derived from marginalizing over a set of variables $L$ and conditioning on $X_S = x_S$ in a set of distributions represented by a directed acyclic graph $\mathcal{D}$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

## 5.2 Chain graphs as unions of DAG models

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

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), p. 4.

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



(i) (ii)

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

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

(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 would be the model obtained by assuming the direction of certain edges being 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 4** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two DAGs with the same adjacencies, such that $\mathcal{D}_{1\cup 2}$ is a chain graph. Then*

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

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

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

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

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

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 $v_1 — \alpha$ or $\alpha — v_2$ (or both).

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

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

This completes the proof. □

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

## 5.3 Ordered blocking of variables

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

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

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

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

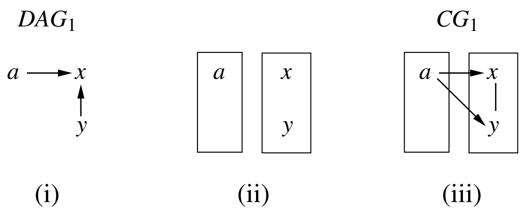



Figure 4: Restricting to chain graph models in keeping with a block ordering may lead to less parsimonious models: (i) $DAG_1$, the generating process; (ii) a block ordering $\{a\} \prec \{x, y\}$; (iii) $CG_1$ the minimal chain graph model for $\{a, x, y\}$, compatible with the ordering which contains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

in Figure 4 (ii) and that, in fact, the simple causal $DAG_1$ in Figure 4 (i) represents the true model.

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

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

In the example just considered there were no unmeasured 'confounding' variables or selection variables.

We now consider the case where such variables may be present. 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 $x_1$, $x_2$ may be ordered relative to $z$, as indicated in the second column of Table 1: $v \sim w$ indicates that $v$ and $w$ are in the same component, while $v \prec w$ indicates that the component containing $v$ precedes the component containing $w$ in the ordering. Note that for (2) and (6) nothing is stated about the relation between the components containing $x_1$ and $x_2$, hence $x_1 \sim x_2$, $x_1 \prec x_2$ and $x_1 \succ x_2$ are all possible in these cases.

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

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

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

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

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

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

In fact it may be shown that any DAG compatible with the ordering given by (6) in which no pairwise marginal independence holds among the variables $x_1$, $x_2$ and $z$, but which satisfies $x_1 \perp\!\!\!\perp x_2 \mid z$, will contain variables

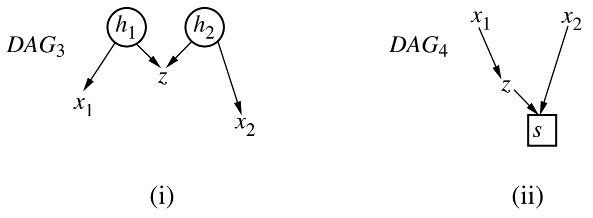



Figure 5: Examples showing that, when latent or selection variables may be present, ordering of variables does not imply the conditional independence relations given by a chain-graph. In $DAG_3$ $h_1$ and $h_2$ are hidden variables. In $DAG_4$ $s$ is a selection variable that has been conditioned on.

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 e.g. Mohamed *et al.* (1998), p. 353.

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

We conclude this section by making several further points:

- The chain graphs in the examples given contained at most three vertices. If we view these graphs as induced subgraphs of a larger chain graph, then the whole discussion carries over if instead of $x_1 \perp\!\!\!\perp x_2 \mid z$, and $x_1 \perp\!\!\!\perp x_2$ we consider $x_1 \perp\!\!\!\perp x_2 \mid W$, with $z \in W$ and $z \notin W$ respectively.
- The problems we have highlighted that arise due to the presence of hidden variables would still be present even if all chain components were singletons, i.e. if we considered DAGs under a fixed ordering.
- There are independence structures arising from DAGs with hidden variables that cannot be represented by any chain graph model. Figure

2 (i) is an example. Wermuth *et al.* (1994, 1999), Koster (1999, 2000), and Richardson and Spirtes (2000) provide 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?

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

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

## 6 Feedback Models for Chain Graphs

As demonstrated by the previous discussion, chain graph models represent qualitatively different hypotheses from those represented by DAG models, including DAG models under marginalization and conditioning. This 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 a number of alternative equilibrium data generating processes with feedback that all lead to chain graph models.

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

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

where $\phi_c$ 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* $\psi_c$ 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 $V = \{1, \dots, p\}$. 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

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

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` $f(y_i \mid x_{-i})$;

`Return` $x$.

The (random) output $X_\tau$ of this program will have distribution $P$ as desired.

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

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

### 6.1.2 The random Gibbs sampler

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

$x \leftarrow x^0$;

`Repeat until equilibrium:`

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

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

`Return` $x$.

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

### 6.1.3 Time reversible Markov dynamics

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

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

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

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

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

### 6.1.4 Langevin diffusions

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

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

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

### 6.1.5 The Gaussian case

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gaussian distribution with mean zero and covariance matrix equal to the unique solution of the matrix equation

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

Clearly there are many more choices for $C$ and $\Lambda$ 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

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

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

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

Since the density obtained by conventional conditioning is given as

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

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

Note that if we consider a more general dynamic regime such as the diffusion (12) this may no longer be true. 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*

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

*if and only if $C$ is symmetric and negative definite. It then holds that $C = -\Sigma^{-1}/2$, where $\Sigma$ 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 (13) is clearly $\Sigma = -C^{-1}/2$. Thus (12) is the Langevin diffusion and the intervention formula (14) holds.

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

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

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

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

and its covariance matrix is the unique symmetric solution $\Omega_{BB}$ to the equation

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

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

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

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

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

and inserting this into (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. $\square$

### 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:

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

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

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

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

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

```
Function G_τ;
    input (x_pa(τ), ε^τ);
```

$x_\tau \leftarrow x_\tau^0$;

$n \leftarrow 0$;

`Repeat until equilibrium:`

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

$n \leftarrow n + 1$;

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

`Return` $x_\tau$.

Here $h_v^\tau$ 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. $h_v^\tau$ is a direct Monte Carlo simulator for this conditional distribution.

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

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

If we order the chain components as $\tau_1, \dots, \tau_p$ and the variables in each chain component $\tau_i = \{n_i + 1, \dots, n_i + t_i\}$ 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` $i = 1, \dots, p$

$j \leftarrow 0$;

`Repeat until equilibrium:`

$j \leftarrow j + 1 \bmod t_i$

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

`Return` $x$;

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

**Proposition 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 assigment system as above, then $X$ has distribution $P$.*

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

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

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

Since equilibrium 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 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 the formula

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

This specializes to the intervention formula (5) in the fully directed case and Bayes' formula in the undirected case: in the fully directed case, all chain components are singletons, so either $\tau \setminus A$ 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

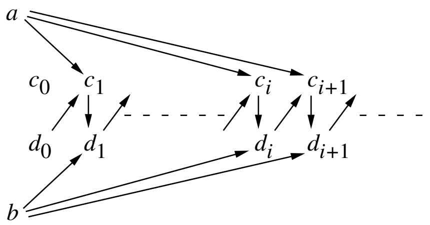



Figure 6: Infinite DAG corresponding to a structural assignment system for $CG_3$ of Figure 1 where $c$ is updated before $d$ in each inner loop.

*et al.* (1999), where interventions are then described by decision nodes. Since

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

where $\tilde{Z}$ is a normalizer as before, an alternative argument for (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 DAGs

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 Figure 1 (ii), the generating process corresponding to this graph using the systematic Gibbs sampler dynamics would first independently choose values $x_a$ and $x_b$ for the variables labelled $a$ and $b$, then use these as input for an equilibrium process updating of $c$ and $d$ as indicated in Figure 6. Using the global Markov property on the DAG in Figure 6 yields

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

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. Thus taking a snapshot as

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

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

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

# 7 Linear Structural Equation Models

## 7.1 Basic terminology

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

There is a unique error term $\epsilon_v$ 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 $\epsilon_v$. In vector notation:

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

where $\gamma_{vv} = 0$. In any given structural model some off-diagonal entries in $\Gamma$ may also be fixed at zero, depending on the form of the structural equations. If under some rearrangement of the rows, $\Gamma$ 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 $\gamma_{vu}$ is not fixed at zero, a SEM is recursive precisely when this graph is a DAG. In a non-recursive system, there might be edges between vertices in both directions if $\gamma_{uv}$ and $\gamma_{vu}$ 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 write (19) as

$$X \leftarrow \Gamma X + \boldsymbol{\epsilon}.$$

If $\Gamma$ is lower triangular and has zero in the diagonals, this expression can be given an unambiguous 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 a SEM model a multivariate normal distribution over the error terms is specified as e.g. $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Delta)$. In any particular model, some off-diagonal ($\delta_{ij}$) entries in $\Delta$ may be allowed to be non-zero. If $\Delta$ is not diagonal then the model is said to have *correlated errors.*

If $(I - \Gamma)$ is non-singular, the traditional interpretation of a SEM determines a joint distribution over the substantive variables by solving the equations (19) to obtain the *reduced form* equations:

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

yielding

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

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 (19). This can make a radical difference, in particular when the effect of interventions are considered. See for example Pearl (1998) and Spirtes *et al.* (1998) for a detailed discussion of these and other issues concerning structural equation models.

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 Figure 2 (iii) with a relabelling of the vertices as $(a, b, c, d) = (1, 2, 3, 4)$:

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

Fisher (1970) presents a dynamic process, whose time average gives the distribution described by a linear non-recursive structural equation model. Here the system is occasionally subjected to random exogeneous disturbances of the exact equilibrium. The eigenvalues of $\Gamma$ 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

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

### 7.2 Chain graph models for structural equations

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

To make such an interpretation we associate an undirected edge with every pair $(u, v)$ for which $\gamma_{uv}$ and $\gamma_{vu}$ are 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 Figure 1 (ii).

The graph of a SEM under this interpretation may not in general be a chain graph and unless this is the case, the model will not have a chain graph interpretation. But if it is, the dynamic schemes discussed in Section 6 could be used to give an alternative interpretation of a structural equation model with feedback.

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 compatible with the chain graph and then each part of the assignment system is interpreted through Gibbs updating as

$$X_\tau(t+1) \xleftarrow{G} \Gamma_\tau * X_\tau(t) + \Gamma_{\tau,\mathrm{pa}(\tau)} * x_{\mathrm{pa}(\tau)} + \epsilon_\tau(t+1)$$

where the subscripted matrices are appropriate submatrices of $\Gamma$ 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

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

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 the variables have been scaled to have unity error variances — 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

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

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

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

If the conditions above are fulfilled, the distribution after intervention as $X_4 \leftarrow x_4^*$ becomes the same as in (20), but now it is obtained from the joint distribution by the intervention formula (18). The joint distribution is different under the chain graph interpretation of the SEM, for which (19) would not lead to the distribution (20).

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

# 8 Discussion

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

## 8.1 Causal DAGs versus causal chain graphs

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 feedback 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 vs. non-recursive SEMs.

## 8.2 Undirected edges and causal under-determination

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

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

Thus only certain undirected edges may be interpreted as (prudently) representing a collection of causal hypotheses; 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 compatible with a pre-specified ordering will often be incompatible with finding the most parsimonious model. This seems undesirable:

- If the primary goal of the analysis is prediction (of the joint distribution) then parsimonious models are often preferable.
- If explanation is the goal then a less parsimonious model — which will include 'extra' edges — may often be misleading, see Figure 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. On the other hand, if one has good reason to believe parametric cancellation is present, then this might argue against attempting to model the independence structure in order to understand the generating process.

The alternative of directly modelling the conditional independence structure without assuming it arises from a generating process appears intractable; even for only four variables there are 18300 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 in order to narrow down a set of candidate models.

### 8.4 Chain graphs under the alternative Markov property

An alternative Markov property for chain graphs has been proposed by Andersson *et al.* (1996). Hence, in general, there are different statistical models that may be associated with the same chain graph. For example, with this alternative interpretation the graph $CG_3$ in Figure 1 (ii) 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 $DAG_4$ in Figure 2 (i). However, there are other chain graphs for which the alternative property results in an independence model that again cannot be obtained from any finite DAG by marginalizing or conditioning (Richardson 1998).

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

### 8.5 Summary

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

## Acknowledgements

This research was supported in part by the Danish Research Councils through their PIFT programme under the project DINA and the U.S. NSF Division of Mathematical Sciences (DMS-9972008). In addition the authors gratefully acknowledge inspiration and support from the ESF 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.

## References

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

Andersson, S. A., Madigan, D., and Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, **25**, 505–41.

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

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

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

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

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

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

Box, G. E. P. (1966). Use and abuse of regression. *Technometrics*, **8**, 625–9.

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

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

Cox, D. R. (1984). Design of experiments and regression. *Journal of the Royal Statistical Society, Series A*, **147**, 306–15.

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

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

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

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

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1–31.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, **95**, 407–48.

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

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

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

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

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

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

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

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

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

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

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

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, **1**, 49–75.

Hofmann, R. (2000). Inference in Markov blanket networks. Technical Report FKI-235-00, Technical University of Munich.

Hofmann, R. and Tresp, V. (1998). Non-linear Markov networks for continuous variables. In *Advances in Neural Information Processing Systems 10. Proceedings of the 1997 Conference*, (ed. M. I. Jordan, M. J. Kearns, and S. A. Solla), pp. 521–7. MIT Press, Cambridge, Mass.

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

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

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

Koster, J. T. A. (1999). Linear structural equations and graphical models. Lecture notes, The Fields Institute, Toronto.

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

Koster, J. T. A. (2000). Marginalizing and conditioning in graphical models. Technical report, Erasmus University, Rotterdam.

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

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

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

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

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

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

Matúš, F. (1999). Conditional independences among four random variables III: Final conclusion. *Combinatorics, Probability and Computing*, **8**, 269–76.

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

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

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

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. In Polish. English translation of Section 9 by D. Dabrowska and T. P. Speed in *Statistical Science* **5** (1990), 465–480.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, **66**, 688–701.

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

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

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

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search.* Springer-Verlag, New York. Reprinted by MIT Press.

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

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

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

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

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

Studený, M. and Bouckaert, R. R. (1998). On chain graph models for description of independence structures. *Annals of Statistics*, **26**, 1434–95.

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

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

Wermuth, N., Cox, D., and Pearl, J. (1994). Explanations for multivariate structures derived from univariate recursive regressions. Technical Report 94-1, University of Mainz, Mainz, Germany.

Wermuth, N., Cox, D., and Pearl, J. (1999). Explanations for multivariate structures derived from univariate recursive regressions. Technical Report Revision of 94-1, University of Mainz, Mainz, Germany.

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

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

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

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

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

## Appendix: Limiting Behaviour of Gibbs Dynamics

In the following we let $q = (q_v)_{v \in V}$ denote a family of conditional specifications, i.e. $q_v(\cdot \mid x_{-v})$ 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 $q_v(\cdot \mid x_{-v})$ is equal to $\mathcal{X}_v$ for all $v \in V$, i.e. that

$$q_v(A \mid x_{-v}) > 0 \text{ for all open sets } A. \tag{22}$$

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

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

If we introduce the transition kernel $Q_v$

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

we may rewrite (23) in a shorter form

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

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

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

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

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 $\pi$. The condition (22) ensures 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 $\sigma$.*

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

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

The converse follows by renumbering $V$. □

Consequently we obtain:

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

(i) $\mu P(\pi) = \mu$ *for all* $\pi \in S(p)$*;*

(ii) $\mu P(\sigma) = \mu$ *for all* $\sigma \in S(p)$ *with* $\sigma$ *cyclic;*

(iii) $\mu = \mu Q_i$ *for all* $i \in V$.

**Proof** We show (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (i). The implication (i) $\Rightarrow$ (ii) is trivial. If (ii) holds, and $i \in V$, we get

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

Thus $\mu Q_i$ 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 (iii).

That (iii) $\Rightarrow$ (i) is easily shown by repeated use of the relations $\mu Q_i = \mu$:

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

This completes the proof. □

Further, we have the following

**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 $\pi$.*

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 $\pi$ in the sitewise updating. If the state space is finite, transient behaviour is not possible.

If the state space is infinite and $|V| \geq 3$, it may be that the Gibbs sampler converges to equilibrium for one permutation $\pi$, but shows transient behaviour for another permutation $\pi'$, provided $\pi$ and $\pi'$ 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 $\mu$, but the specifications $q$ will not be conditional distributions with respect to this $\mu$ 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 the latter reference has subsequently 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 $\mu$ of a stationary pseudo-Gibbs sampler and the conditional specifications $q$.