Relating graphical frameworks; Undirected, directed acyclic and chain graph models

Christopher Meek[1]

Department of Philosophy
Carnegie Mellon University,
Pittsburgh, PA 15213
cm1x@andrew.cmu.edu

Abstract

Researchers have systematically explored the Markov equivalence relationships among models from three classes of graphical models; the undirected, the directed acyclic, and the chain graph or block-recursive models. This paper considers the Markov equivalence relationships between models of different classes of graphical models and gives conditions for the existence of a model from a given class which is "almost Markov equivalent" to a given model from another class. An example of such a condition which is well known in the literature is the Wermuth condition. In addition, for each of the classes of models correct algorithms for learning models from conditional independence facts are given. While correct algorithms for learning undirected and directed acyclic graphs from only conditional independence facts exist in the literature, no such algorithms exists for chain graphs. This paper presents algorithms for learning undirected and directed acyclic graphs to highlight the relationship between the learning problems for these classes of models and the learning problem for chain graphs. The learning algorithms are proved correct and are shown to run in polynomial time in the number of vertices for fixed degree graphs except for one algorithm for learning undirected graphs which is polynomial in the number of vertices regardless of the degree of the graph.

[1]Research for this paper was supported by the Office of Navel Research grant ONR #N00014-93-1-0568.

1 Introduction

In this paper several classes of graphical models are compared and algorithms for learning models in each class are given. The classes are the undirected, directed acyclic and chain graph models. The directed acyclic graphical models have a long history in statistical modeling (see Sewall Wright 1921) while the undirected and chain graph models more recent innovations. Accounts of recent work in statistical modeling and decision-making under uncertainty using directed graphical representations can be found in Pearl (1988), Schacter (1986) and Spirtes et al. (1993). This work goes under many names including influence diagrams, belief networks, and Bayesian networks. An account of recent work on the use of undirected graphs can be found in Whittaker (1990) and Pearl (1988). The most recent of these classes of models are the chain graphs which were introduced in Lauritzen and Wermuth (1989) and further developed in Frydenberg (1990). Chain graphs are a class of models which essentially subsume both the undirected and directed acyclic graphical models. While correct algorithms for learning undirected and directed acyclic graphs exist in the literature no such algorithm exists for chain graphs. This paper remedies this deficiency and gives additional results which relate each of these classes of models.

2 Notation and definitions

A *graph* is a pair $G = \langle V, E\rangle$ where V is a finite set of vertices (which correspond to random variables) and E is a set of edges, i.e. a subset of the set of all ordered pairs of distinct vertices, $Ord(V) = \{\langle\alpha,\beta\rangle | \alpha \in V \wedge \beta \in V \wedge \alpha \neq \beta\}$. We write (and draw) $\alpha \to \beta$ if and only if $\langle\alpha,\beta\rangle \in E \wedge \langle\beta,\alpha\rangle \notin E$ and $\alpha — \beta$ if and only if $\langle\alpha,\beta\rangle \in E \wedge \langle\beta,\alpha\rangle \in E$. When we draw a graph there is no edge between two vertices if and only if neither ordered tuple is in the set of edges. α is adjacent to β if and only if $\alpha — \beta$, $\alpha \to \beta$, or $\beta \to \alpha$; we use $\alpha \in ADJ(\beta)$ to denote this. A graph is *undirected* if and only if for no pair of vertices α and β is it the case that $\alpha \to \beta$. A graph is *directed* if and only if for no pair of vertices α and β is it the case that $\alpha — \beta$.

An ordered n-tuple $\langle\alpha_1, \ldots, \alpha_n\rangle$ $(n > 1)$ of distinct vertices is called a *path from* α_1 *to* α_n *in graph* G if and only if for $1 \leq i < n$ it is the case $\langle\alpha_i, \alpha_{i+1}\rangle \in E$. A path is *directed* if and only if for some i it is that case

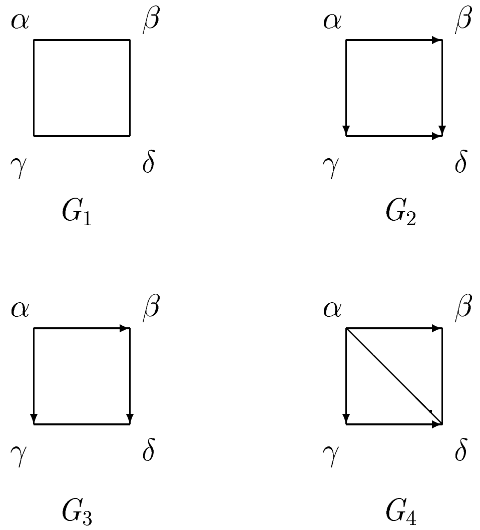


Figure 1: Example graphs

that $\alpha_i \rightarrow \alpha_{i+1}$ and a path is *undirected* if and only if for all i it is the case that $\alpha_i - \alpha_{i+1}$. $\alpha < \beta$ if and only if there is a directed path from α to β. The *descendants* of a vertex α are those vertices in $de(\alpha) = \{\beta | \alpha < \beta\}$. The *nondescendants* of a vertex α are those vertices in $nd(\alpha) = V \backslash de(\alpha)$. An n-tuple $\langle \alpha_1, \ldots, \alpha_{n-1}, \alpha_1 \rangle$ is a *cycle* if and only if $\langle \alpha_1, \ldots, \alpha_{n-1} \rangle$ and $\langle \alpha_{n-1}, \alpha_1 \rangle$ are paths. A graph is *acyclic* if and only if there is no directed cycle (i.e. no edge in any cycle is directed).

If A is a subset of the vertices then the *induced subgraph* given A is $G_A = \langle A, E_A \rangle$ where $E_A = E \cap Ord(A)$. The *underlying undirected graph* of a graph G is $G^u = \langle V, E^u \rangle$ where $E^u = \{\langle \alpha, \beta \rangle | \langle \alpha, \beta \rangle \in E \vee \langle \beta, \alpha \rangle \in E\}$. The *boundary* of vertex α in graph G is the set of vertices connected to α by a directed or undirected edge; $bd(\alpha) = \{\beta | \beta \rightarrow \alpha \vee \beta - \alpha\}$. More generally the definition of the function bd applied to sets of vertices is defined as $bd(A) = \cup_{\beta \in A} bd(\beta)$. The example graphs in Figure 1 and Table 2 illustrate some of the definitions.

Let P be a probability measure over $\mathcal{X}_V$ (a space for the random variables in V) and $G = \langle V, E \rangle$ be an acyclic graph. The pair $\langle P, G \rangle$ satisfy the *local Markov condition* with respect to acyclic graph G if and only if for all $\alpha \in V$ it is the case that $\alpha \perp\!\!\!\perp (nd(\alpha) \backslash (bd(\alpha) \cup \{\alpha\})) | bd(\alpha)$. Every vertex is independent of its nonboundary nondescendants given its boundary. $Markov(G)$ is the

	$de(\beta)$	$nd(\beta)$	$bd(\delta)$	acyclic
G_1	$\emptyset$	$\{\alpha, beta, \gamma, \delta\}$	$\{\beta, \gamma\}$	yes
G_2	$\{\delta\}$	$\{\alpha, \beta, \gamma\}$	$\{\beta, \gamma\}$	yes
G_3	$\{\gamma, \delta\}$	$\{\alpha, \beta\}$	$\{\beta, \gamma\}$	yes
G_4	$\{\gamma\}$	$\{\alpha, \beta, \delta\}$	$\{\alpha, \beta, \gamma\}$	no

Table 1: Examples

set of distributions that satisfy the local Markov condition with respect to G. Two graphs, G and G' are Markov equivalent if and only if $Markov(G) = Markov(G')$.

A distribution P over random variables V satisfies the intersection property if and only if for all disjoint subsets X, Y, W, and Z of V it is the case that $X \perp\!\!\!\perp Y|ZW \wedge X \perp\!\!\!\perp W|YZ \Rightarrow X \perp\!\!\!\perp YW|Z$. The intersection property holds for positive distributions. $Markov^+(G)$ is the set of distributions that satisfy the local Markov condition with respect to G and the intersection property. Two graphs, G and G' are *almost Markov equivalent* if and only if $Markov^+(G) = Markov^+(G')$.

Let $\tau(\alpha)$ be the set of all vertices reachable from α by an undirected path. For a directed graph $\tau(\alpha) = \{\alpha\}$. The *moral graph* of $G = \langle V, E\rangle$ is the graph $G^m = \langle V, E^m\rangle$ where $E^m = E^u \cup (\bigcup_{\alpha \in V} Ord(bd(\tau(\alpha))\backslash\tau(\alpha)))$. In the case where the graph is directed and acyclic the moralization of the graph corresponds to connecting all pairs of parents for each vertex in the graph and then undirecting all of the edges. In an undirected graph, S *separates* A and B if and only if every path from a member of A to a member of B contains a member of S. The *anterior set* of A with respect to an acyclic graph G is the smallest set containing (i) every member of A (ii) every member in $\cup_{\beta \in A}\tau(\beta)$ and (iii) every vertex $\alpha < \beta$ for some $\beta \in \cup_{\gamma \in A}\tau(\gamma)$. Let $an(A)$ denote the anterior set for A.

Pearl (1988) and Lauritzen et al. (1990) have given rules that can be used to infer independence facts about probability distributions in $Markov(G)$ and $Markov^+(G)$ from the graphical structure of a directed acyclic graph G. The two rule are equivalent in the case of directed acyclic graphs but Lauritzen's rule generalizes to arbitrary acyclic graphs.

Definition 1 *(Lauritzen's rule)* $G \vdash A \perp\!\!\!\perp B|S$ *if and only if* S *separates* A *and* B *in* $(G_{an(A \cup B \cup S)})^m$

Definition 2 $P \models A \perp\!\!\!\perp B|S$ *if and only if the independence fact* $A \perp\!\!\!\perp B|S$ *holds in the joint distribution* P.

A set of inference rules is sound for a class of models if and only if the statements derivable by the rules are true of all models in the class of models. A set of inference rules is complete for a class of models if and only if all of the statements true in all models are derivable by the rules. In the cases that considered in this paper the statements are of the form $A \perp\!\!\!\perp B|S$ and the sets of models are sets of distributions which satisfy the local Markov condition with respect to some graph and possibly the intersection property. The importance of the completeness of a set of inference rules is that one can be insured that no further facts can be correctly inferred from the graphical structure and the local Markov condition. The pair $\langle G, P \rangle$ satisfy the *faithfulness* condition if and only if the only independence facts true in P are exactly those entailed by Lauritzen's rule; i.e. $P \models A \perp\!\!\!\perp B|S$ if and only if $G \vdash A \perp\!\!\!\perp B|S$. A class of distributions $\mathcal{P}$ has the *strong completeness* property if and only if for all G there exists a distribution $P \in \mathcal{P}$ such that $\langle G, P \rangle$ satisfies the faithfulness condition. In this paper the class of distribution will be the class of all probability distribution; see Meek (1995b) for a more detailed discussion of strong completeness, faithfulness and completeness with respect to classes of distributions of interest. Clearly strong-completeness entails completeness. The relationship between the local Markov condition and Lauritzen's rule and completeness is examined for each of the classes of graphs.

3 Undirected graphical models (ugs)

In this section we present several results about undirected graphical models. For more details about undirected graphical models see Whittaker (1990), Pearl (1988) and Frydenberg (1990). We begin by investigating the usefulness of Lauritzen's rule as an inference rule.

Theorem 1 (Soundness; Pearl, Lauritzen) *Let* G *be an undirected graph. If* S *separates* A *and* B *in* $(G_{an(A \cup B \cup S)})^m$ *then* $A \perp\!\!\!\perp B|S$ *in every distribution in* $Markov^+(G)$.

Theorem 2 (Strong Completeness; Geiger et al.) *For all undirected graphs G there is a probability distribution such that the pair $\langle G,P \rangle$ satisfy the local Markov condition, the intersection property and such that all and only those independence facts which follow from Lauritzen's rule are true.*

Theorem 3 *Two undirected graphs $G = \langle V, E \rangle$ and $G' = \langle V', E' \rangle$ are Markov equivalent if and only if $V = V'$ and $E = E'$.*

Proof — Suppose the two graphs are identical. Then the independence constraints imposed by the Markov condition are identical and thus the set of Markov distributions are identical. The converse follows from Theorem 2 and the fact that two non-identical graphs over V have different adjacencies and thus different separation properties.□

3.1 Learning undirected graphs

Several authors have considered the problem of learning undirected probabilistic models including Fung and Crawford (1990), and Pearl (1988). In this section I present two correct algorithms for inducing probabilistic models and compare the two algorithms.

If find-ug (or find-ug2) is given a complete graph on n variables and a faithful joint distribution P over those variables then find-ug (or find-ug2) will find the smallest undirected graph G such that $\langle G, P \rangle$ satisfy the local Markov condition, i.e. it will find the undirected graph to which the distribution is faithful.

```
Function maxdegree(G:graph):integer;
;; this function returns the maximal number of adjacencies for any
;; vertex in G
   begin
      maxdegree = 0
      for each vertex A in G do
         if sizeof(ADJ(A)) > maxdegree then maxdegree = sizeof(ADJ(A))
      return(maxdegree)
   end
```

```
Function find-ug(G:graph;P:distribution;Sep:array of sets):graph;
;; Sep is an n x n array of sets of vertices
;; which is used in the next section
   begin
      n=0
      repeat
         for each ordered pair of adjacent vertices A, B do
            begin
              for each subset S of ADJ(A)\{B} of size n do
                 if A is independent of B given S
                 then
                    begin
                        remove the edge between A and B from G
                        Sep(A,B) = S
                    end

            end
         n = n + 1
      until maxdegree(G) < n
      return(G)
   end

Function find-ug2(G:graph;P:distribution):graph;
   begin
      for each ordered pair of vertices A, B do
         if A is independent of B given V\{A,B}
         then remove the edge between A and B from G
   end
```

The correctness of these algorithms rests upon the correctness of the statistical tests which are performed and the assumption of faithfulness. The correctness of the two algorithms follows from the following argument. Let G be the unique undirected graph to which P is faithful. Two vertices α and β which are adjacent in G are not separated in G by any set S and thus will remain connected in the final graph. If the two vertices α and β are not adjacent then there is some set S which separates the two vertices. All that needs to be argued is that this separating set S will be found. In the case of find-ug2 this is trivial. In the case of the find-ug algorithm simply notice that α is separated from β in G by the boundary of α (i.e. $bd(\alpha)$) and the algorithm does not terminate until every independence fact corresponding to a separating set of at least the size of $maxdegree(G) \geq bd(\alpha)$ is checked.

Algorithms for performing tests of conditional independence are not given but see Bishop et al. (1975) or Fienberg (1977) for tests in multinomial distributions and Whittaker (1990) for tests in multivariate normal distributions. The complexity analysis of find-ug and find-ug2 is straight forward if we assume that the complexity of testing conditional independence is constant; this assumption is reasonable in many contexts. Let n be the number of vertices in the graph and k be the maximum degree of the graph to be learned. In the case of find-ug2 the complexity of the algorithm is $O(n^2)$ since there are that many pairs of vertices in the graph. In the case of find-ug the complexity is $O(n^{k+2})$ since, in the worst case, the algorithm requires $2\begin{pmatrix} n \\ 2 \end{pmatrix} \sum_{i=0}^{k} \begin{pmatrix} n-2 \\ i \end{pmatrix}$ conditional independence tests. Although the number of independence tests used in the find-ug2 algorithm is often less than the number used in find-ug, find-ug2 is not practical in many cases because it uses less powerful tests which are computationally less tractable, this is especially true when using discrete data. In addition to its statistical and computational advantages, the find-ug algorithm will be used as the basis of learning algorithms in later sections.

4 Directed acyclic graphs (dags)

See Pearl (1988) and Lauritzen et al. (1990) for additional properties of directed acyclic graphs.

Theorem 4 (Soundness; Pearl, Lauritzen) *Let G be a directed acyclic graph. If S separates A and B in $(G_{an(A \cup B \cup S)})^m$ then $A \perp\!\!\!\perp B | S$ in every distribution in Markov(G).*

Theorem 5 (Strong Completeness; Geiger et al.) *For all directed acyclic graphs G there is a probability distribution such that the pair $\langle G,P \rangle$ satisfy the local Markov condition, the intersection property and all and only those independence facts which follow from Lauritzen's rule are true.*

The *pattern* for the directed graph G is the graph which has the identical adjacencies as G and which has an oriented edge $\alpha \rightarrow \beta$ if and only if there is a vertex $\gamma \notin ADJ(\alpha)$ such that $\alpha \rightarrow \beta$ and $\gamma \rightarrow \beta$ in G. Let $pattern(G)$ denote the pattern for G. A triple $\langle \alpha, \beta, \gamma \rangle$ of vertices is an *unshielded collider* in G if and only if $\alpha \rightarrow \beta$, $\gamma \rightarrow \beta$ and α is not adjacent to β. It is easy to show that two directed acyclic graphs have the same pattern if and only if they have the same adjacencies and same unshielded colliders.

Theorem 6 (Pearl 1988) *Two directed acyclic graphs, G and G' are Markov equivalent if and only if $pattern(G) = pattern(G')$.*

Two graphs with the same pattern have the same Markov entailed independence facts; this follows from the fact that two graphs have the same moralized anterior graph for every triple of disjoint sets of variables. The converse follows from Theorem 5.

4.1 Learning directed graphs

Given the relationship between independence and graphical structure when on assumes the Markov condition, several authors (Pearl 1988 and Spirtes et al. 1993) have used statistical tests of independence to guide the selection of directed acyclic models. As in the case of undirected graphs, reliable learning algorithms exist for learning the graphical structure for a distribution P if P is faithful to some directed acyclic graph.[2] If find-dag is given a complete graph on n variables and a faithful joint distribution P over those variables then find-dag will find the pattern of a directed graph G to which the distribution P is faithful.

```
Function find-pattern(G:graph; P:distribution; Sep:array of sets):graph;
   begin
      for each unshielded triple <A,B,C> do
         begin
```

[2]Many other approaches to learning directed acyclic models exists inluding Bayesian (see Heckerman et al. 1994 and Cooper et al. 1992), minimum description length and information theoretic methods (see Sclove 1994). Each of these alternative approaches may be termed scoring methods since scores are given to models and model selection is the process of finding the model which maximizes or minimizes a given score criterion. In this paper only independence approaches are discussed.

```
                if B is not in Sep(A,C)
                then orient <A,B> and <C,B>
          end
   end

Procedure find-dag(G:graph; P:distribution);
   begin
      Sep = n x n array of empty sets
      G = find-ug(G,P,Sep)
      G = find-pattern(G,P,Sep)
      return(G)
   end
```

The find-dag algorithm is essentially the same as the PC algorithm given in Spirtes et al. (1993). The algorithm does not return a directed acyclic graph but rather the pattern of any graph in a specific Markov equivalence class. Let us assume that graph G in Figure 2 is a graph to which P is faithful. The PC algorithm would output the graph $pattern(G)$ also shown in Figure 2. If the desired output is a directed acyclic graph in the Markov equivalence class represented by $pattern(G)$ then the algorithm described in Meek (1995) can be used to extend $pattern(G)$ to such a directed acyclic graph.

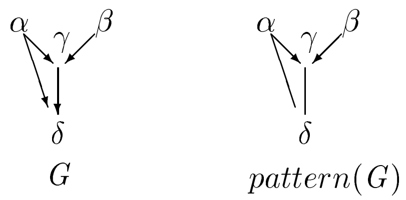


Figure 2: Graph and pattern

The reliability of the PC algorithm rests upon the correctness of statistical tests and upon the assumption that P is faithful to some directed acyclic graph. Let G be the graph to which P is faithful. The correctness of the first step of PC (i.e. find-ug) follows essentially as in the correctness of the find-ug algorithm applied to undirected graphs; basically, any pair of vertices which are adjacent in G will remain adjacent and for any pair of vertices which are not adjacent in G a set will be found which separates the pair of vertices. The correctness of the second part of the algorithm (i.e. find-pattern) follows from the fact that for an unshielded collider $\langle \alpha, \beta, \gamma \rangle$ in G it is the case that α and γ are dependent conditional upon any set containing β. If $\langle \alpha, \beta, \gamma \rangle$ is an unshielded noncollider then α and γ are dependent conditional upon any set not containing β. These facts are readily apparent if one considers the moralized anterior graphs for the two cases. The algorithm simply finds a set S upon which α and γ are conditionally independent and check whether β is in S and orients the unshielded triple accordingly.

The complexity of the find-pattern algorithm is $O(n^3)$. Thus for graphs with maximum degrees larger than one the complexity of the find-ug procedure dominates and thus, in such cases, the complexity of find-dag is $O(n^{k+2})$. It is important to note that the complexity of the learning procedure can be significantly reduced if one has structural background knowledge. For instance, if one has a total ordering on the variables the learning problem is $O(n^2)$ for arbitrary degree graphs.

Note that the correctness of the algorithm rest upon slightly weaker assumption than the faithfulness of P; only the tests which are actually used in the algorithm are required to be correct.

4.2 Relating directed and undirected graphs

A directed acyclic graph satisfies the *Wermuth condition* if and only if there are no unshielded colliders in G.

Theorem 7 *A directed acyclic graph G has an almost Markov equivalent undirected graph if and only if G satisfies the Wermuth condition. The undirected Markov equivalent graph for G is G^u.*

Proof — If directed acyclic graph $G = \langle V, E \rangle$ satisfies the Wermuth condition then for all $B \subseteq V$ it is the case that $G^m_B = G^u_B$. Thus if α and β are separated by S in $G^m_{an(\{\alpha,\beta\} \cup S)}$ if and only if they are separated in G^u and thus they have the same Markov entailed independence facts.

In the case where directed acyclic graph G does not satisfy the Wermuth condition there is an unshielded collider $\langle \alpha, \beta, \gamma \rangle$. Let P_G be a distribution which is faithful to G and satisfies the intersection property; one exists by

Theorem 5. We know that $G \vdash \alpha \perp\!\!\!\perp \gamma | S$ for some set $S \subseteq V \backslash \{\alpha, \beta, \gamma\}$ and that for all sets $S' \subseteq V \backslash \{\alpha, \beta, \gamma\}$ it is not the case that $P_G \models \alpha \perp\!\!\!\perp \gamma | S' \cup \{\beta\}$. Suppose there is an almost Markov equivalent undirected graph H. Let P_H be a distribution which is faithful to H and satisfies the intersection property; one exists by Theorem 2.

(Case i) H has an edge between α and γ. We know that for all S it is not the case that $P_H \models \alpha \perp\!\!\!\perp \gamma | S$ thus $P_H \notin Markov^+(G)$, a contradiction.

(Case ii) H does not have an edge between α and γ. For some set S'' it is the case that $H \vdash \alpha \perp\!\!\!\perp \gamma | S'' \cup \{\beta\}$. But it is not the case that $P_G \models \alpha \perp\!\!\!\perp \gamma | S'' \cup \{\beta\}$ thus $P_G \notin Markov^+(H)$, a contradiction.□

It is easy to write an algorithm to test if there exists a Markov equivalent undirected graph for a given directed acyclic graph exists and to find such an undirected graph if one exists.

A cycle $\langle \alpha_1, \ldots, \alpha_n, \alpha_1 \rangle$ is *chordal* if and only if for some non-consecutive pair of vertices α_i and α_j $(1 \leq i \leq n-2$ and $i+2 \leq j \leq n)$ α_i is adjacent to α_j. A graph is said to be a *chordal graph* if and only if every cycle is chordal.

Theorem 8 *An undirected graph G has an almost Markov equivalent directed acyclic model if and only if G is a chordal graph.*

Proof — A total ordering $\prec$ on the vertices of an undirected graph G induces a directed acyclic graph $G_\prec$ by the rule that if $\alpha — \beta$ in G then orient as $\alpha \to \beta$ in $G_\prec$ if $\alpha \prec \beta$. A total order $\prec$ is *consistent* for undirected graph G if and only if $G_\prec$ satisfies the Wermuth condition. An undirected graph has a consistent total ordering if and only if it is chordal (see Meek 1995). By Theorem 7, an undirected graph has an almost Markov equivalent directed acyclic graph if and only if it is chordal.□

Tarjan and Yannakakis (1984) have given a linear time algorithm to check the chordality of a graph. An algorithm for finding a directed acyclic graph G' which is Markov equivalent to a chordal undirected graph G can be found in Meek (1995).

5 Chain Graphs

A graph G is a *chain graph* if and only if G is acyclic. Every directed acyclic and every undirected graph is a chain graph.

Theorem 9 (Soundness; Frydenberg) *Let G be an acyclic graph. If S separates A and B in $(G_{an(A \cup B \cup S)})^m$ then $A \perp\!\!\!\perp B | S$ in every distribution in $Markov^+(G)$.*

In the case of directed and undirected graphs, strong completeness results have been proven but strong completeness result for the case of the chain graphs has been published although several authors have conjectured that such a result holds (e.g. Frydenberg 1990).

A triple $\langle \alpha, B, \gamma \rangle$ is a *complex* if and only if (i) $B \subseteq \tau(\delta)$ for some $\delta \in B$, (ii) $\alpha \in bd(B)$ and $\gamma \in bd(B)$ and (iii) $\alpha \notin ADJ(\gamma)$. A triple $\langle \alpha, B, \gamma \rangle$ is a *minimal complex* if and only if (i) the triple is a complex, and (ii) there is no $B' \subseteq B$ such that $\langle \alpha, B', \gamma \rangle$ is a complex.

We extend the notion of pattern to handle chain graphs. The pattern of a chain graph G is the graph (i) with the same adjacencies as G and (ii) the edge $\alpha — \beta$ is oriented $\alpha \rightarrow \beta$ if and only if there exists a vertex γ and a set B containing β such that $\langle \alpha, B, \gamma \rangle$ is a minimal complex.

Theorem 10 (Frydenberg) *Two chain graphs $G = \langle V, E \rangle$ and $G' = \langle V', E' \rangle$ are almost Markov equivalent if and only if $pattern(G) = pattern(G')$.*

5.1 Learning chain graphs

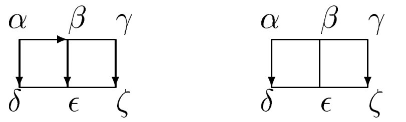
As in the case of undirected and directed graphs, reliable learning algorithms exist for learning the graphical structure for a distribution P if P is faithful to some chain graph. If find-cg is given a complete graph on n variables and a faithful joint distribution P over those variables then find-cg will find the pattern of a graph to which P is faithful. Alternative methods for inducing chain graphs have been published but these require large amounts of background assumptions and strong modeling assumptions. For instance, Højsgaard and Thiesson (1992) require that the user specifies the block structure[3] and that the models within the blocks are decomposable (see Pearl 1988).

[3] In the notation of this paper, the user must specify disjoint sets of variables $\langle A_1, \ldots, A_n \rangle$ such that for all $1 \leq i \leq n$ and for all $\alpha \in A_i$ it is the case that $\tau(\alpha) = A_i$ and $\cup_i A_i = V$.

```
Procedure orient-cg(G,P)
   begin
      for all pair of vertices  A, B such that A not in ADJ(B) do
         begin
            find maximal subset SA of ADJ(A) such that
               A is independent of B given SA
            find maximal subset SB of ADJ(B) such that
               A is independent of B given SB
            if C in ADJ(A) and C not in SA then orient <A,C>
            if C in ADJ(B) and C not in SB then orient <A,C>
         end
   end

Function Find-cg(G,P);
   begin
      Sep = n x n array of empty sets
      G = find-ug(G,P,Sep)
      G = orient-cg(G,P)
      return(G)
   end
```

As in the case of learning directed acyclic graphs, the output of find-cg is a pattern which is not necessarily a chain graph. A chain graph G and its corresponding pattern $pattern(G)$ are shown in Figure 3. If P is a distribution which is faithful to the chain graph G then the output of the find-cg algorithm would be $pattern(G)$. Note that $pattern(G)$ is not a chain graph since $\langle \alpha, \delta, \epsilon, \beta, \alpha \rangle$ and $\langle \gamma, \zeta, \epsilon, \beta, \gamma \rangle$ are directed cycles.



G $pattern(G)$

Figure 3: Chain graph and pattern

The correctness of the find-cg algorithm is similar to that of the PC algorithm.[4] The correctness of the first step (i.e. find-ug) is completely analogous and will not be repeated. To show the correctness of the second step (i.e. orient-cg) we need the following three facts.

Fact 1 *Let G be a chain graph and α and γ be a pair of vertices not adjacent in G. If there exists a B such that $\langle \alpha, B, \gamma \rangle$ is a minimal complex and $\beta \in B$ is such that $\beta \in ADJ(\alpha)$ then for all sets $S \subseteq V$ which contains at least one member of B it is not the case that $G \vdash \alpha \perp\!\!\!\perp \gamma | S$.*

Fact 1 follows trivially from Lauritzen's rule.

Fact 2 *For all chain graphs G and pairs of vertices α and γ not adjacent in G there exists a unique maximum (written $maximum(\alpha, \gamma)$) set S such that (i) $S \subseteq ADJ(\alpha)$ and (ii) $G \vdash \alpha \perp\!\!\!\perp \gamma | S$.*

Proof — Suppose not. Then there exists maximal sets $S_1 \subseteq ADJ(\alpha)$ and $S_2 \subseteq ADJ(\alpha)$ such that $S_1 \Delta S_2 \neq \emptyset$ and $G \vdash \alpha \perp\!\!\!\perp \gamma | S_1$ and $G \vdash \alpha \perp\!\!\!\perp \gamma | S_2$.[5] Let β be a vertex in $S_1 \Delta S_2$ and with out loss of generality $\beta \in S_1$. A vertex δ is *connected* to a vertex ϵ if and only if (i) $\delta \in \tau(\epsilon)$ or (ii) $\delta < \epsilon$ or (iii) $\epsilon < \delta$. From the fact that $G \vdash \alpha \perp\!\!\!\perp \gamma | S_1$ and $G \vdash \alpha \perp\!\!\!\perp \gamma | S_2$ it must be the case that β is not connected to γ. Thus $G \vdash \alpha \perp\!\!\!\perp \gamma | S_2 \cup \{\beta\}$ and we have a contradiction.$\square$

Fact 3 *Let G be a chain graph and α and γ be a pair of vertices not adjacent in G and $\beta \in ADJ(\alpha)$. Let $maximum(\alpha, \gamma)$ be the set described in Fact 2; the maximum subset of $ADJ(\alpha)$ which makes α and γ independent. If $\beta \to \alpha$ or $\beta — \alpha$ then $\beta \in maximum(\alpha, \gamma)$.*

Fact 3 follows from Lauritzen's rule and the observation that $\beta \in an(\{\alpha\})$.

Fact 2 guarantees that we can find the maximum sets SA and SB.Fact 1 guarantees that if α is involved in a minimal complex $\langle \alpha, B, \gamma \rangle$ then the vertex $\beta \in B$ such that $\beta \in ADJ(\alpha)$ will not be in $maximum(\alpha, \gamma)$ and thus

[4]Several modifications can be made to make the orient-cg algorithm more efficient but these would needlessly complicate the presentation.

[5]$S_1 \Delta S_2$ is the symmetric difference of the two sets.

the edge will be oriented correctly. Finally Fact 3 guarantees that for each $\beta \in ADJ(\alpha)$ such that $\beta \rightarrow \alpha$ or $\beta — \alpha$ then the edge will not be oriented $\alpha \rightarrow \beta$ in the output of find-cg. Since every pair of nonadjacent vertices is checked and the respective maximal sets for each of the pair is found every edge involved in a minimal complex will be found. Thus the algorithm is shown to be correct.

The complexity of the orient-cg algorithm is $O(n^3)$. Thus for graphs with maximum degrees larger than one the complexity of the find-ug procedure dominates and thus, in such cases, the complexity of find-cg is $O(n^{k+2})$. As in the case of learning directed acyclic models, structural background knowledge can significantly improve the efficiency of learning algorithms.

The correctness of the algorithm rests upon the assumption that the distribution P is faithful to some chain graph. The difficulty with assuming faithfulness in this context is that there is no proof that there exists a faithful distribution exists for an arbitrary chain graph. One response to such a criticism is to conjecture that such a faithful distribution exists for any arbitrary chain graph out of an analogy to the undirected and directed cases. Such a conjecture, while not unsound, would not be a reasonable answer to such worries. A second response is that the reliability of the procedure does not rest upon the full strength of the assumption of faithfulness (as in the undirected and directed cases). But again the existence of a distribution for this restricted notion of faithfulness has not been demonstrated. A final response is that even in the event of failures of faithfulness procedures based upon the assumption (e.g. the PC algorithm) have proved to be a useful basis of learning directed acyclic graphs (see Spirtes unpublished). While none of these responses is completely satisfying only time (and empirical study) will tell if such a procedure is useful.

5.2 Relating directed and undirected graphs and chain graphs

Not all patterns of directed acyclic graphs are chain graphs. Figure 4 gives example of a directed acyclic graph whose pattern is not a chain graph; the pattern fails to be a chain graph because of the directed cycle $\langle \alpha, \gamma, \delta, \alpha \rangle$.

A pattern is a graphical representation of an entire class of Markov equivalent graphs. This example shows that the pattern of a graph can not be used

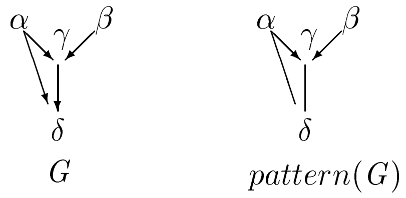


Figure 4: Graph and pattern

as a chain graph representation of the the set of Markov equivalent graphs. However, a canonical graph of a directed graph G (defined below) is a chain graph and represents the entire set of graphs which are Markov equivalent to G. First, in words, a *canonical graph* for directed acyclic graph G (written $cg(G)$) is the graph which represents all of the adjacencies and orientation common to graphs which are Markov equivalent to G.

$$\begin{gathered} cg(G) = \langle V, E' \rangle \text{ where} \\ E' = \bigcup \{ E' | G' = \langle V, E' \rangle \wedge pattern(G) = pattern(G') \} \end{gathered}$$

Theorem 11 (Meek;Madigan) *For all directed acyclic graphs G cg(G) is a chain graph*[6]

Theorem 12 *For all directed acyclic graphs G, G and $cg(G)$ are almost Markov equivalent.*

Theorem 13 *A chain graph G has an almost Markov equivalent undirected graph if and only if G has no minimal complexes.*

Theorem 14 *A chain graph G has an almost Markov equivalent directed graph if and only if G has no non-chordal undirected cycle and each of the minimal complexes in G are unshielded colliders (i.e. if $\langle \alpha, B, \gamma \rangle$ is a minimal complex then B is a singleton set).*

Theorem 12, Theorem 13, and Theorem 14 follow from Theorem 10.

[6]In Meek (1995) the concept of $cg(G)$ is equivalent to the concept of $max(pattern(G), \emptyset)$. Madigan's result is as of yet unpublished.

6 Final Remarks

In summary this paper presents algorithms for learning undirected, directed acyclic, and chain graph models. The paper also contains additional results about the existence of almost Markov equivalent models in a given class for a model in a second class. There are several avenues for continued research.

- As noted in the previous section, the pattern Π of a chain graph is not necessarily a chain graph. While it is easy to write inefficient algorithms to find a chain graph G with the same minimal complexes as the pattern (i.e. $pattern(G) = \Pi$) no efficient algorithms for this task have been published except for the special case where the pattern is the pattern of a directed acyclic graph (see Meek 1995). A closely related question originally posed by Frydenberg (1990) is the problem of finding the chain graph with the same minimal complexes as a given chain graph and with the smallest number of directed edges possible. Frydenberg has shown that there is a unique model of this description for any given chain graph.

- Two closely related open questions are the whether Lauritzen's rule is strongly complete for chain graphs and whether there exists a faithful distribution for arbitrary chain graphs.

- An empirical and theoretical evaluation of the reasonableness of faithfulness as an inferential tool for model selection, in particular, learning chain graphs.

References

Bishop, Y., S. Fienberg, and P. Holland (1975). *Discrete Multivariate Analysis.* Cambridge, Mass.: MIT Press.

Cooper, G. and E. Herskovits (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning 9.*

Fienberg, S. (1977). *The analysis of cross-classified categorical data.* Cambridge, Mass: MIT Press.

Frydenberg, M. (1990). The chain graph markov property. *Scand. J. Statistics 17.*

Geiger, D. (1990). *Graphoids: A qualitative Framework for Probabilistic Inference.* Ph. D. thesis, UCLA.

Heckerman, D., D. Geiger, and D. Chickering (1994). Learning bayesian networks: The combination of knowledge and statistical data. Technical Report MSR-TR-94-09, Microsoft Research.

Højsgaard, S. and T. B. (1992). Bifrost - block recursive models induced from relevant knowledge, observations, and statistical techniques. Technical report, Institute of Electronic Systems, Aalborg, Denmark.

Lauritzen, S., A. Dawid, B. Larsen, and H. Leimer (1990). Independence properties of directed markov fields. *Networks 20.*

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

Meek, C. (1995). Causal inference and causal explanation with background knowledge. Technical report, Philosophy Department, Carnegie Mellon University.

Meek, C. (1995b). Strong-completeness and faithfulness in belief networks. Technical Report CMU-PHIL-62, Philosophy Department, Carnegie Mellon University.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent systems.* San Mateo: Morgan-Kaufmann.

Schacter, R. (1986). Evaluating influence diagrams. *Operations Research 34*(6).

Sclove, S. (1994). Small-sample and large-sample statistical model selection criteria. In *Selecting Models from Data*, pp. 31–41. Springer-Verlag.

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

Whittaker, J. (1990). *Graphical Models in applied multivariate statistics.* Wiley.

Wright, S. (1921). Correlations and causation. *Journal of Agricultural Research 20.*