Parameterizing and Scoring Mixed Ancestral Graphs

Thomas Richardson

&

Peter Spirtes

August 1999

Technical Report No. CMU-PHIL-102

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Parameterizing and Scoring Mixed Ancestral Graphs

10/11/98 10:06 PM

by Thomas Richardson[1] (tsr@stat.washington.edu) and

Peter Spirtes (ps7z@andrew.cmu.edu) and

---

[1] The first author was a Rosenbaum Fellow at the Isaac Newton Institute during the preparation of this paper, and wishes to thank the Institute for all the support provided.

1) Introduction

In recent years, directed acyclic graphs (DAGs) have received considerable attention. (Good overviews can be found in Lauritzen 1996, Pearl 1988, and Whittaker 1990.) DAGs represents a set of distributions, all of which share certain conditional independence relations encoded in the graph. As long as a DAG does not contain any latent variables, and any sample contains data that is missing at random, DAGs have a number of attractive features that have made them widely used in both statistics and Artificial Intelligence. However, introducing latent variables and data that is not missing at random makes the estimation, evaluation, and finding of DAG models considerably more difficult. In this paper we will discuss a class of graphical models, the mixed ancestral graphs, which are a generalization of DAGs, are closed under marginalization and conditionalization, retain the attractive features of other graphical models, and can handle latent variables and data that is not missing at random while making the estimation, evaluation, and search problems more tractable.

First we will review the definition of MAGs and state some of their more important features. Then we will show how to form a linear parameterization of MAGs, how to perform maximum likelihood estimates of the parameters, demonstrate that the set of distributions represented by a linear MAG is a curved exponential family, show how to calculate the dimensionality of a linear MAG, and how to calculate the BIC score of a linear MAG. Finally we will illustrate the use of MAGs on an actual data set, and illustrate the advantages that they have over other kinds of graphical models.

2) DAGs and MAGs

Each missing edge in a DAG entails some conditional independence relation. In the case of DAGs, the conditional independence relation entailed by a missing edge is a "local Markov" condition: if there is no edge between X and Y, and Y is not an ancestor

of X, then X is independent of Y given the parents of Y.[2] (Pearl 1988, Lauritzen *et al.* 1990) There is a three place graphical relation among disjoint subsets of variables in a DAG G (**X** is d-separated from **Y** given **Z**) that holds if and only if satisfying the local Markov condition for G entails that **X** is independent of **Y** given **Z**; this relation is described in more detail in section ??? (Pearl 1988, Lauritzen *et al.* 1990)

A DAG can also be given a simple interpretation as a data-generating mechanism (in which the value of a variable X is generated by the values of X's parents.) (Wermuth and Lauritzen 1990??, Spirtes *et al*, 1993.) When interpreted in this way, DAGs are useful for predicting the effects of changing an existing data-generating mechanism. (Spirtes *et al.* 1993, Pearl 1995)

Often a DAG G contains latent variables, and selection variables (which take the value 1 if a unit is in the sample, and 0 otherwise.)[3] Suppose there is a distribution P(**V**) represented by DAG G. Assume that the variables in **V** can be partitioned into **O** (observed), **L** (latent), and **S** (selected, or conditioned on.) (**S** variables are used in Rubin ??) In DAGs with latent variables and selection bias, we place latent variables (which are marginalized out) in boxes and selection variables (which are conditioned on) in circles. In that case instead of observing P(**V**), we may be able to observe only P(**O**|**S** = **1**), that is the marginal distribution over the observed variables in the selected subpopulation. Let us call P(**O**|**S** = **1**) the "observed" distribution. A DAG G in which the variables have been partitioned into observed variables, latent variable, and selection variables will be denoted by G(**O**,**S**,**L**).

A MAG is a graph that represents both the d-separation relations among the variables in **O** (condition on the variables in **S**), and represents some of the ancestor

---

[2] This follows from the more general rule that each variable is independent of its non-parental non-descendants given its parents.

[3] In general, corresponding to each variable $O_n \in \mathbf{O}$, there is a variable $S_n \in \mathbf{S}$, which takes the value 1 for a unit in which O_1 has been recorded, and 0 otherwise. If there are no missing values for observed variables in the sample, then there is a single selection variable S, which is 1 if the unit is in the sample, and 0 otherwise.

relations among the variables in **O** or **S**. In order to define MAGs, we will first define MGs (mixed graphs), and m-separation (a generalization of d-separation.)

3) Definition of MAGs

A **mixed graph (MG)** is a graph <**V**,**E**> with three kinds of edges in **E**, **directed edges** A → B (also written as B ← A); **double-headed edges** A ↔ B; **and undirected edges** A — B. At most one edge connects any given pair of vertices. An example of an MG that is not a DAG, an UG, or a chain graph is shown in Figure 1. We define the following graph theoretical notions in an MG G with vertices **V**, which are simple generalizations of the corresponding concepts in DAGs and undirected graphs.

A sequence of vertices $<V_1,...,V_{n+1}>$ is a **path** if for $1 \bullet i \bullet n$, there is an edge with endpoints V_i and V_{i+1}, and $E_i \neq E_{i+1}$. A path U is **acyclic** if no vertex appears more than once in the corresponding sequence of vertices. We will assume that a path is acyclic unless specifically mentioned otherwise. A sequence of vertices $<V_1,...,V_{n+1}>$ is a **directed path** D **from** V_1 to V_{n+1} if and only if for $1 \bullet i \bullet n$, there is a directed edge $V_i \rightarrow V_{i+1}$. U is a **m-directed path** (abbreviating mixed graph directed path) from X to Y if there is a path U between X and Y such that if A and B are adjacent on U, and A is between X and B or X = A, then the edge between A and B is out of A (e.g. <B,D,A,E> in M in Figure 1. A is a **parent** of B (and B is a **child** of A) in a MAG M if there is an edge A → B; A is an **ancestor** of B (and B is a descendant of A) in M if there is a directed path from A to B or A = B. X is an **m-ancestor** of Y in MAG G if there is a m-directed path from X to Y or X = Y (e.g. B is a m-ancestor of E in M in Figure 1.) $X \in$ **Ancest(R)** if and only if X is an ancestor of some member of **R**. $X \in$ **M-Ancestors(R)** if X is a m-ancestor of a member of **R**. If a MAG M is a DAG, then X is a m-ancestor of Y if and only if it is an ancestor of Y. V is a **collider** on a path P in MAG M if P contains two adjacent edges that are into V (e.g. E is a collider on <D,A,E,F> in M.) An **u-vertex** (undirected vertex) in a MAG M is a variable A for which there is some edge A — B.

4) M-Separation

D-separation is a graphical relationship in DAGs that is useful for determining whether an arbitrary conditional independence relation is entailed by satisfying the local

Markov condition for a DAG. For disjoint sets of vertices **X**, **Y**, and **Z** in DAG M, **X** is d-connected to **Y** if there is a path U between $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ such that every collider on U is an ancestor of **Z**, and no non-collider on U is in **Z**; otherwise **X** is **d-separated** from **Y** given **Z**. If **X** and **Y** are d-separated given **Z** in DAG G, then for any distribution that satisifies the local Markov condition for G, **X** and **Y** are independent conditional on **Z**. (See Lauritzen *et al.* 1990 for an exposition on the relation between various separation principles.)

M-separation is a generalization of d-separation that plays a role analogous to that of d-separation. The exact role that m-separation plays in MGs will be explained in more detail in section ??. The definition of m-separation and m-connection in MGs carries over unchanged from the definition of d-separation and d-connection respectively in DAGs. (Of course, terms such as "collider" in the definition of m-separation are generalizations of the corresponding term in the definition of d-separation.) This entails that m-separation (m-connection) when applied to a DAG is identical to d-separation (d-connection).

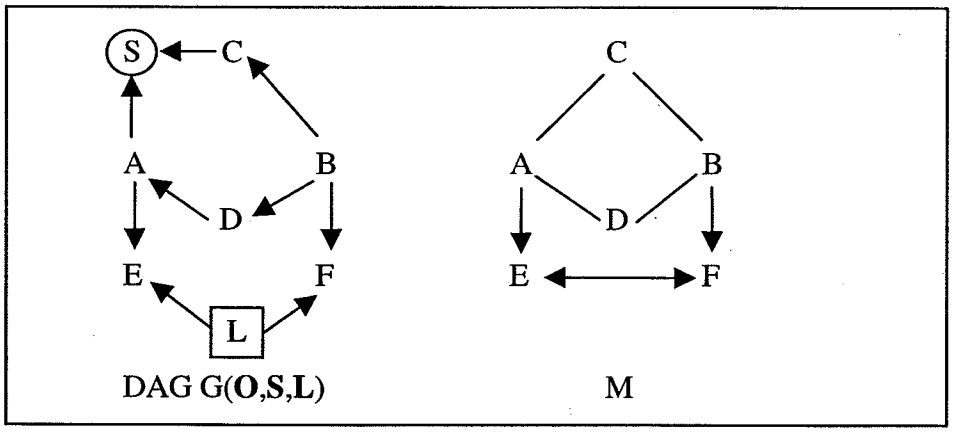


Figure 1

5) Mixed Ancestral Graphs (MAGs)

We will be concerned with a subclass of MG's called Mixed Ancestral Graphs (or MAGs). An MG M is a **MAG** if:

1. If for every subset **W** of **O**, X_i and X_j are m-connected given $\mathbf{W} \cup \mathbf{S}$, then X_i and X_j are adjacent in M.

2. If X_i and X_j are adjacent in M, and X_i is an ancestor of X_j in M, then the edge between X_i and X_j is oriented as $X_i \rightarrow X_j$.
3. If X_i is an u-vertex in M, then every edge containing X_i is out of X_i.

Condition 1 is required in order to ensure that if X and Y are not adjacent, then there is some set that m-separates them.

Just as the vertices in a DAG can be partitioned into <**O**,**S**,**L**>, so the vertices in a MAG M can be partitioned into <**O**,**S**,**L**>; in that case we write M(**O**,**S**,**L**). If we refer to a MAG M with vertices **V**, then it is assumed that the partition of **V** is <**V**,∅,∅>.

A MAG M with vertices **V** is said to **represent** another MAG G(**O**,**S**,**L**) when **V** = **O**, and for **X**, **Y**, **Z** disjoint subsets of **O**, **X** is m-separated from **Y** given **Z** ∪ **S** in G(**O**,**S**,**L**) if and only if **X** is m-separated from **Y** given **Z** in M. If a MAG M represents a DAG G(**O**,**S**,**L**), then m-separation in M gives the conditional independence consequences of P(**O** ∪ **S** ∪ **L**) satisfying the local Markov condition for G(**O**,**S**,**L**) in P(**O**|**S**=1).

An example of a MAG is shown in Figure 1, where **O** = {A,B,C,D,E,F}, members of **L** are enclosed in ovals, and members of **S** are enclosed in boxes and M represents G(**O**,**S**,**L**).

The following algorithm describes how to form a MAG that represents both the d-separation relations and some of the ancestor relations among the variables in **O** and **S** in MAG M(**O**,**S**,**L**).

Algorithm Form-MAG

Input: MAG G(**O**,**S**,**L**)

Output: MC(G(**O**,**S**,**L**))

1. Place the edge A — B in MC(G(**O**,**S**,**L**)) if and only if A is an m-ancestor of B or **S** in G(**O**,**S**,**L**), B is an m-ancestor of A or **S** in G, and in G for every subset **W** of **O**, A and B are m-connected given **W** ∪ **S**.

2. Place the edge A → B in MC(G(**O**,**S**,**L**)) if and only if A is an m-ancestor of B or **S** in G(**O**,**S**,**L**), B is not an m-ancestor of A and **S** in G, and and in G for every subset **W** of **O**, A and B are m-connected given **W** ∪ **S**.

3. Place the edge A $\leftrightarrow$ B in MC(G(**O**,**S**,**L**)) if and only if A is not an m-ancestor of B and **S** in G, B is not an m-ancestor of A and **S** in G, and and in G for every subset **W** of **O**, A and B are m-connected given **W** $\cup$ **S**.

It is clear from the algorithm that MC(G(**O**,**S**,**L**)) represents som of the m-ancestor relations in G(**O**,**S**,**L**). (If G(**O**,**S**,**L**) is a DAG, then the m-ancestor relations are just ancestor relations.) The next theorem states that the result of applying this operation to a MAG M is another MAG which represents M.

Theorem 1: If for some MAG G(**O**,**S**,**L**), M = MC(G(**O**,**S**,**L**)), then M is a MAG that represents G(**O**,**S**,**L**).

For example, for G(**O**,**S**,**L**) and M in Figure 1, M = MC(G(**O**,**S**,**L**)).

6) Parameterizing MAGs

A **linear parameterization of a MAG M** is a model parameterized in the following way:

- Two u-vertices A and B have a non-zero correlation conditional on all of the other u-vertices only if there is an edge A — B in M.
- Each non u-vertex A in M is a linear function of its parents in M, and a unique error term, ε_A.
- Two error terms ε_A and ε_B have a non-zero correlation only if there is an edge A $\leftrightarrow$ B in the graph.
- Each u-vertex is uncorrelated with each error term.

For notational convenience, we will assume that the variables in a MAG G are $X_1,...,X_s,X_{s+1}...X_n$, where X_i is not an ancestor of X_j in G if i > j, s is the number of u-vertices in G, and all of the u-vertices precede all of the non u-vertices. We will refer to the error term of X_i as ε_i, rather than ε_{X_i}.

A **complete MAG** is a MAG in which every pair of variables is adjacent.

Lemma 1: If G is a MAG, there is a complete MAG G_C, such that G is a subgraph of G_C.

Proof. Suppose that G is a MAG. Form G_C in the following way: if X_i and X_j are u-vertices then add an edge X_i — X_j; if X_i is an u-vertex and X_j is an non u-vertex then add an edge $X_i \rightarrow X_j$; if X_i and X_j are n-variables and X_i is an ancestor of X_j in G, add an edge $X_i \rightarrow X_j$ in G_C; and if X_i and X_j are n-variables and X_i is not an ancestor of X_j and X_j is not an ancestor of X_i, then add an edge $X_i \leftrightarrow X_j$ to G_C. By Theorem 1, G_C is a MAG. ∴

In MAG M, $V \in$ **Ancest(X)** if $V \notin \mathbf{X}$, and V is an ancestor of some vertex $X \in \mathbf{X}$, or V is an u-vertex .

Theorem 2: If G_C is a complete MAG over a set of variables X, and Σ is a positive definite covariance matrix for X, then there is a linear parameterization θ of G_C such that $\Sigma_{G_C(\theta)} = \Sigma$.

Proof. Let Σ be the covariance matrix for **X**. An instantiation of a parameterization of G_C has the properties that each non u-vertex can be expressed as linear function of its parents and an error term, that if $\text{cov}(\varepsilon_p,\varepsilon_q) \bullet 0$ then $X_p \leftrightarrow X_q$ in G_C, and if X_p — X_q then the partial correlation of X_p and X_q conditional on the other u-vertices • 0. We will now show that there is a parameterization of G_C that has covariance matrix Σ. Let **Parents**(X_k) be the set of parents of X_k.

Note that since G_C is a complete ancestral graph, if X_k is an non u-vertex, then **Parents**(X_k) $\subseteq \{X_j | j < k\}$, and further if $X_i \in \{X_j | j < k\} \setminus$ **Parents**(X_k) then $X_i \leftrightarrow X_k$ in G_C. We will abbreviate **Parents**(X_k) by $\mathbf{P}_k$. Take each variable X_k in turn. Regress X_k on $\mathbf{P}_k$. Let

$$\hat{X}_k = \sum_{X_j \in \mathbf{P}_k} \alpha_{kj} X_j$$

be the linear predictor of X_k on $\mathbf{P}_k$ (where summation over an empty set is equal to zero) and the residuals

$$\varepsilon_k = X_k - \sum_{X_j \in \mathbf{P}_k} \alpha_{kj} X_j$$

We will now show that the α_{kj} and the correlations between the residuals form a parameterization of the complete MAG G_C. First note that X_k is a linear function of its parents in G_C and the error term ε_j because

$$X_k = \sum_{X_j \in \mathbf{P}_k} \alpha_{kj} X_j + \varepsilon_j$$

Second, we will show that if $\text{Cov}(\varepsilon_p,\varepsilon_q) \neq 0$ then $X_p \leftrightarrow X_q$ in G_C, (where X_p and X_q are non u-vertices, because they have error terms). Suppose that $\text{Cov}(\varepsilon_p, \varepsilon_q) \neq 0$, but that there is no double headed arrow $X_p \leftrightarrow X_q$ in G_C. We may suppose without loss of generality that $p < q$. Since there is no double headed arrow $X_p \leftrightarrow X_q$, and $p < q$ it follows that $X_p \rightarrow X_q$ in G_C. It then follows that $X_p \in \mathbf{P}_q$.

$$\text{Cov}(\varepsilon_q,\varepsilon_p) = \text{cov}(\varepsilon_q, X_p - \sum_{X_j \in \mathbf{P}_p} \alpha_{pj} X_j)$$

We will now show that $\text{cov}(\varepsilon_q,\varepsilon_p) = 0$ by showing that $\text{cov}(\varepsilon_q,X_p) = 0$, and for all X_j in $\mathbf{P}_p$, $\text{cov}(\varepsilon_q,X_j) = 0$. By construction, ε_q is uncorrelated with $X_p \in \mathbf{P}_q$, (since ε_q is the residual remaining after regressing X_q on $\mathbf{P}_q$), so $\text{Cov}(\varepsilon_q,X_p) = 0$. If $X_j \in \mathbf{P}_p$, then $X_j \rightarrow X_p$ in G_C. Since $X_p \rightarrow X_q$ in G_C, it follows that X_j is an ancestor of X_q in G_C. As G_C is a complete ancestral graph it then follows that $X_j \rightarrow X_q$ in G_C, so $X_j \in \mathbf{P}_q$. Hence $\text{cov}(\varepsilon_q,X_j) = 0$, as claimed. It follows that $\text{cov}(\varepsilon_q,\varepsilon_p) = 0$.

If X_p is an u-vertex, and X_q is an non u-vertex, then by construction, ε_q is uncorrelated with $X_p \in \mathbf{P}_q$, (since ε_q is the residual remaining after regressing X_q on $\mathbf{P}_q$), so $\text{cov}(\varepsilon_q,X_p) = 0$.

Finally, positive definiteness of Σ ensures that each ε_k has positive variance; otherwise $X_k$would be a linear combination of previous X_i's and Σ would not be positive definite. $\therefore$

Lemma 2: In a MAG G, if $X_j \in$ **Ancest**(X_i), and there is no edge $X_j \rightarrow X_i$ and X_i is not an non u-vertex, then X_i is m-separated from X_j given **Ancest**$(X_i)\backslash\{X_j\}$.

Proof. Suppose, on the contrary that there is a path U that m-connects some member $X_j \in$ **Ancest**(X_i) to X_i given **Ancest**$(X_i)\backslash\{X_j\}$. There are four cases: on U either there is an edge $X_k \rightarrow X_i$, an edge $X_i \rightarrow X_k$, an edge $X_i \leftrightarrow X_k$, or an edge $X_k - X_i$. Because any u-vertex on a path is a non-collider, and every u-vertex (except possibly X_j) is in **Ancest**$(X_i)\backslash\{X_j\}$, U does not contain any u-vertex (except possibly X_j), and hence U does not contain $X_k - X_i$.

Suppose there is an edge $X_k \rightarrow X_i$ on U. $X_k \neq X_j$ because otherwise there is an edge $X_j \rightarrow X_i$ in G. Hence X_k is in **Ancest**$(X_i)\backslash\{X_j\}$. But then X_k is not a collider on U, and U does not m-connect X_i to X_j given **Ancest**$(X_i)\backslash\{X_j\}$.

Suppose that the first edge on U is an edge $X_i \rightarrow X_k$. It follows that either X_i is an ancestor of X_j, or there is a collider on U. Because G is acyclic, and X_j is an ancestor of X_i, X_i is not an ancestor of X_j. Suppose then that there is a collider on U. Let X_l be the first collider on U; it follows that X_i is an ancestor of X_l. Because U m-connects X_i and X_j given **Ancest**$(X_i)\backslash\{X_j\}$, X_l is an ancestor of **Ancest**$(X_i)\backslash\{X_j\}$. Because X_l is not an u-vertex but is in **Ancest**$(X_i)\backslash\{X_j\}$, X_l is an ancestor of X_i. It follows that G is cyclic, contrary to our assumption that G is a MAG.

Suppose that the first edge on U is an edge $X_i \leftrightarrow X_k$. It follows that either X_k is an ancestor of X_j, or there is a collider on U. If X_j is an u-vertex, then X_k is not an ancestor of X_j (because no u-vertex has any non-trivial ancestors.) If X_j is not an u-vertex then it is an ancestor of X_i, in which case X_k is an ancestor of X_i, contrary to the $X_i \leftrightarrow X_k$ edge. Suppose then that there is a collider X_l on U. Because U m-connects X_i and X_j given **Ancest**$(X_i)\backslash\{X_j\}$, X_l is an ancestor of X_i. It follows that G is cyclic, contrary to our assumption that G is a MAG.∴

U is an **inducing path** between X and Y with respect to MAG M(**O**,**S**,**L**) if and only U is an acyclic path such that every member of $\mathbf{O} \cup \mathbf{S}$ on U is a collider on U, and every collider on U is an ancestor of $\{X,Y\} \cup \mathbf{S}$. This is a generalization of the concept of inducing path that was introduced in Verma and Pearl 1990. The following two lemmas are proved in Richardson and Spirtes ??

Lemma 3: In MAG G(**O**,**S**,**L**) if there is an inducing path U between A and B then for any subset **Z** of $\mathbf{O}\backslash\{A,B\}$ there is a path P that m-connects A and B given $\mathbf{Z} \cup \mathbf{S}$ with the same orientation as U.

Lemma 4: In MAG G(**O**,**S**,**L**) if a path U m-connects A and B given ((**M-Ancestors**$(\{A,B\} \cup \mathbf{S}) \cap \mathbf{O}) \cup \mathbf{S})\backslash\{A,B\}$ then U is an inducing path.

Theorem 3: If G is a MAG, and Σ is a positive definite covariance matrix such that if X_i and X_j are m-separated given **Z** in G, then $\text{cov}(X_i,X_j|\mathbf{Z}) = 0$, then there is a linear parameterization θ of G such that $\Sigma_{G(\theta)} = \Sigma$.

Proof. By Lemma 1, there is a complete MAG G_C such that G is a subgraph of G_C. By Theorem 2 there is a parameterization θ of G_C such that $\Sigma_{G_C(\theta)} = \Sigma$. We will now show that θ assigns zeroes to every edge that is in G_C but not in G.

First consider an edge X_i — X_j that is in G_C but not in G. Because there is no edge between X_i and X_j in G, every path between X_i and X_j contains some vertex that is an u-vertex and not a collider. **Ancest**(X_i,X_j) is equal to the set of u-vertices. Hence X_i and X_j are m-separated given **Ancest**(X_i,X_j). Hence $\text{cov}_\Sigma(X_i,X_j|\textbf{Ancest}(X_j,X_i)) = 0$ by hypothesis, and the parameter in G_C associated with the X_i — X_j edge is 0.

Next consider an edge $X_i \rightarrow X_j$ that is in G_C but not in G. By the method of construction of G_C, X_i is in **Ancest**$(X_j)\backslash\{X_i\}$ in G. Because G does not contain $X_i \rightarrow X_j$, by Lemma 2, X_j is m-separated from X_i given **Ancest**$(X_j)\backslash\{X_i\}$ in G. Hence $\text{cov}_\Sigma(X_i,X_j|\textbf{Ancest}(X_j)\backslash\{X_i\}) = 0$ by hypothesis.

Because G_C is a MAG, ε_j is uncorrelated with the errors of any member of **Ancest**$(X_j)\backslash\{X_i\}$, and hence uncorrelated with any member of **Ancest**$(X_j)\backslash\{X_i\}$. Hence in θ_C the coefficients of the ancestors of X_j in the equation for X_j are equal to the partial regression coefficients of X_j on **Ancest**$(X_j)\backslash\{X_i\}$. But when X_j is regressed on **Ancest**$(X_j)\backslash\{X_i\}$, the partial regression coefficient of X_i in the equation for X_j, is equal to zero when $\text{Cov}_{G_C(\theta_C)}(X_i,X_j|\textbf{Ancest}(X_j)\backslash\{X_i\}) = 0$. Hence, if there is no edge $X_i \rightarrow X_j$ in G, X_i and X_j are m-separated given **Ancest**$(X_j)\backslash\{X_i\}$ in G, and by hypothesis $\text{Cov}_\Sigma(X_i,X_j|\textbf{Ancest}(X_j)\backslash\{X_i\}) = 0$. $\text{Cov}_{G_C(\theta_C)}(X_i,X_j|\textbf{Ancest}(X_j)\backslash\{X_i\}) = \text{Cov}_\Sigma(X_i,X_j|\textbf{Ancest}(X_j)\backslash\{X_i\}) = 0$. Hence in θ_C, the partial regression coefficient of X_i in the equation for X_j, is equal to zero.

Finally consider an edge $X_i \leftrightarrow X_j$ that is in G_C but not in G. Let θ be a parameterization of G such that every parameter that is in both θ and θ_C is equal (i.e. θ is the same as θ_C except that it sets parameters corresponding to edges in G_C but not in G equal to zero.)

Let <i,j> be an ordered pair such that $i < j$, and <k,m> an ordered pair such that $k < m$. Say <k,m> < <i,j> if $m < j$ or $m = j$ and $k < i$. Let the induction hypothesis be that if <k,m> < <i,j> then $\text{cov}_{G_C(\theta_C)}(\varepsilon_k,\varepsilon_m) = \text{cov}_{G(\theta)}(\varepsilon_k,\varepsilon_m)$. We will then show that $\text{cov}_{G_C(\theta_C)}(\varepsilon_i,\varepsilon_j) = \text{cov}_{G(\theta)}(\varepsilon_i,\varepsilon_j)$. (This is true by definition of θ if the edge $X_i \leftrightarrow X_j$ occurs in G; we will show that it is true even when $X_i \leftrightarrow X_j$ does not occur in G.)

By Theorem 1 there is no inducing path between X_i and X_j in G. By Lemma 4, X_i and X_j are m-separated given **Ancest**(X_iX_j) in G. Hence $\mathrm{cov}_{G(\theta)}(X_i,X_j|\mathbf{Ancest}(X_iX_j)) = 0$. By hypothesis, $\mathrm{cov}_{G_C(\theta_C)}(X_i,X_j|\mathbf{Ancest}(X_iX_j)) = \mathrm{cov}_{\Sigma}(X_i,X_j|\ \mathbf{Ancest}(X_iX_j)) = 0$. Hence,

(1) $$\mathrm{cov}_{G(\theta)}(X_i,X_j|\mathbf{Ancest}(X_iX_j)) = \mathrm{cov}_{G_C(\theta_C)}(X_i,X_j|\mathbf{Ancest}(X_iX_j)).$$

First we will show that if $\langle k,m\rangle < \langle i,j\rangle$ then $\mathrm{cov}_{G(\theta)}(X_k,X_m) = \mathrm{cov}_{GC(\theta C)}(X_k,X_m)$.

$$X_k = \sum_{1\le r\le s} a_{kr}X_r + \sum_{s<r\le k} a_{kr}\varepsilon_r \quad X_m = \sum_{1\le r\le s} a_{mr}X_r + \sum_{s<r\le m} a_{mr}\varepsilon_r$$

$$\mathrm{cov}_{G(\theta)}(X_k,X_m) =$$

$$\sum_{1\le r\le \min(k,s)} a_{kr}b_{mr}\,\mathrm{var}_{G(\theta)}(X_r) + \sum_{s<r\le k} a_{kr}b_{mr}\,\mathrm{var}_{G(\theta)}(\varepsilon_r) +$$

$$2\sum_{1\le r\le \min(k,s)}\ \sum_{r<t\le \min(m,s)} (a_{kr}b_{mt} + a_{kt}b_{mr})\,\mathrm{cov}_{G(\theta)}(X_r,X_t)$$

$$2\sum_{s\le r\le k}\ \sum_{r<t\le m} (a_{kr}b_{mt} + a_{kt}b_{mr})\,\mathrm{cov}_{G(\theta)}(\varepsilon_r,\varepsilon_t)$$

Note that in $G_C(\theta_C)$ and $G(\theta)$, the coefficients of the reduced form of each variable is exactly the same in each of the parameterizations, because we have already shown that the structural equations in each parameterization are identical. (The reduced form expresses each variable as a linear function of error variables.) We have already shown that if X_r and X_t are u-vertices, then $\mathrm{cov}_{G(\theta)}(X_r,X_t) = \mathrm{cov}_{G_C(\theta_C)}(X_r,X_t)$. Because $\langle r,t\rangle < \langle i,j\rangle$, by the induction hypothesis $\mathrm{cov}_{G(\theta)}(\varepsilon_r,\varepsilon_t) = \mathrm{cov}_{G_C(\theta_C)}(\varepsilon_r,\varepsilon_t)$. It follows that $\mathrm{cov}_{G(\theta)}(X_k,X_m) = \mathrm{cov}_{G_C(\theta_C)}(X_k,X_m)$. Because $\mathrm{cov}_{G(\theta)}(X_k,X_m) = \mathrm{cov}_{G_C(\theta_C)}(X_k,X_m)$, equations (2), (3), and (4) follow.

(2) $$\mathrm{cov}_{G(\theta)}(X_i, \mathbf{Ancest}(X_i, X_j)) = \mathrm{cov}_{G_C(\theta_C)}(X_i, \mathbf{Ancest}(X_i, X_j))$$

(3) $$\mathrm{var}_{G(\theta)}^{-1}(\mathbf{Ancest}(X_i, X_j)) = \mathrm{var}_{G_C(\theta_C)}^{-1}(\mathbf{Ancest}(X_i, X_j))$$

(4) $$\mathrm{cov}_{G(\theta)}(X_j, \mathbf{Ancest}(X_i, X_j)) = \mathrm{cov}_{G_C(\theta_C)}(X_j, \mathbf{Ancest}(X_i, X_j))$$

By rearranging the terms in $\mathrm{cov}(X_i,X_j|\mathbf{Ancest}(X_i, X_j))$, equations (5) and (6) follow.

(5) $$\mathrm{cov}_{G(\theta)}(X_i,X_j) =$$
$$\mathrm{cov}_{G(\theta)}(X_i,X_j|\mathbf{Ancest}(X_i, X_j)) - \mathrm{cov}_{G(\theta)}(X_i, \mathbf{Ancest}(X_i, X_j)) \times$$
$$\mathrm{var}_{G(\theta)}^{-1}(\mathbf{Ancest}(X_i, X_j)) \times \mathrm{cov}_{G(\theta)}(X_j, \mathbf{Ancest}(X_i, X_j))$$

(6) $$\mathrm{cov}_{G_C(\theta_C)}(X_i,X_j) =$$

$$\mathrm{cov}_{G_C(\theta_C)}(X_i,X_j|\mathbf{Ancest}(X_i,X_j)) - \mathrm{cov}_{G_C(\theta_C)}(X_i,\mathbf{Ancest}(X_i,X_j)) \times$$
$$\mathrm{var}_{G_C(\theta_C)}{}^{-1}(\mathbf{Ancest}(X_i,X_j)) \times \mathrm{cov}_{G_C(\theta_C)}(X_j,\mathbf{Ancest}(X_i,X_j))$$

From equations (1) – (6) it follows that $\mathrm{cov}_{G(\theta)}(X_i,X_j) = \mathrm{cov}_{G_C(\theta_C)}(X_i,X_j)$.

$$X_i = \sum_{1\le r\le s} a_{ir}X_r + \sum_{s<r\le i} a_{ir}\varepsilon_r \qquad X_j = \sum_{1\le r\le s} a_{jr}X_r + \sum_{s<r\le j} a_{jr}\varepsilon_r$$

$$\mathrm{cov}_{G(\theta)}(X_i,X_j) =$$

$$\sum_{1\le r\le \min(i,s)} a_{ir}b_{jr}\,\mathrm{var}_{G(\theta)}(X_r) + \sum_{s<r\le i} a_{ir}b_{jr}\,\mathrm{var}_{G(\theta)}(\varepsilon_r) +$$

$$2\sum_{1\le r\le \min(i,s)}\ \sum_{r<t\le \min(j,s)} (a_{ir}b_{jt}+a_{it}b_{jr})\mathrm{cov}_{G(\theta)}(X_r,X_t)$$

$$2\sum_{s\le r\le i}\ \sum_{r<t\le j} (a_{ir}b_{jt}+a_{it}b_{jr})\mathrm{cov}_{G(\theta)}(\varepsilon_r,\varepsilon_t)$$

An analogous equation holds for $\mathrm{cov}_{G_C(\theta_C)}(X_i,X_j)$. All of the terms in the two equations have been proved equal except $\mathrm{cov}_{G_C(\theta_C)}(\varepsilon_i,\varepsilon_j)$ and $\mathrm{cov}_{G(\theta)}(\varepsilon_i,\varepsilon_j)$. Because $\mathrm{cov}_{G(\theta)}(X_i,X_j) = \mathrm{cov}_{G_C(\theta_C)}(X_i,X_j)$, it follows that $0 = \mathrm{cov}_{G(\theta)}(\varepsilon_i,\varepsilon_j) = \mathrm{cov}_{G_C(\theta_C)}(\varepsilon_i,\varepsilon_j)$. ∴

7) Maximum Likelihood Estimates

The vertices V in a MAG M can be divided into two disjoint subsets, U, consisting of the u-vertices, and N, consisting of all other vertices. Similarly, the parameters θ in a MAG M can be divided into two parts: θ_u, which are the parameters associated with undirected edges, and θ_n, all of the other parameters. The implied covariance matrix among U depends only θ_u, while the implied covariance matrix among N depends upon θ_u and θ_n. This implies that when calculating maximizing maximum likelihood estimates of θ, we can first maximize the likelihood for θ_u, and then for θ_n.

For normal distributions, the maximum likelihood estimates can be found by minimizing the following fit function:

$$F_{ML} = \log|\Sigma_{G(\theta)}| + \mathrm{tr}(S\Sigma_{G(\theta)}{}^{-1}) - \log|S| - p$$

where S is the sample covariance matrix, $G(\theta)$ is the vector of parameters, and p is the number of vertices.

8) Curved Exponential Families and the Dimensionality Of MAGs

Let the set of natural parameters of a regular exponential family be denoted by N.

Theorem 4: The family of distributions represented by a linear MAG M over a set of k variables is a locally parameterized curved exponential family of dimension equal to k(k+1)/2 minus the number of pairs of variables in M that are not adjacent to each other.

Proof. According to Theorem 4.2.1 in Kass and Vos(1997), a subfamily S_0 of an n-dimensional regular exponential family is a locally parameterized curved exponential family if for each η_0 in N_0 there is an open neighborhood U in N containing η_0 and a diffeomorphism h: $U \rightarrow R^k \times R^{n-k}$ such that $S_0^U = \{P_\eta \text{ in } S^U: h(\eta) = (\beta,\psi) \text{ and } \psi = 0\}$.

First we will show that there is a diffeomorphism from a covariance matrix Σ of the normal distribution (with zero means) over k variables to the parameters of a complete MAG. By Corollary A.3 in Kass and Vos, it suffices to show that there is a smooth one-to-one function from Σ to the parameters of a complete MAG, whose inverse is also smooth.

According to Theorem 2, the distributions represented by a given MAG M can be parameterized in the following way. For a given covariance matrix Σ among the $\mathbf{X}$ variables, regress each non u-vertex X_k on the set $\mathbf{P_k} := \{X_j \mid X_j \bullet X_k \text{ and } X_j \in \text{Ancest}(X_k)\}$. Let

$$\hat{X}_k = \sum_{X_j \in \mathbf{P_k}} \alpha_{kj} X_j$$

be the linear predictor of X_k on $\mathbf{P_k}$, or 0 if $\mathbf{P_k}$ is empty. Now let

$$\varepsilon_k = X_k - \sum_{X_j \in \mathbf{P_k}} \alpha_{kj} X_j$$

The α_{kj}, the non-zero covariances among the ε_k, and the partial correlations among the u-vertices parameterize a MAG. It follows that

$$\text{cov}(\varepsilon_p,\varepsilon_q) = \text{cov}(X_p - \hat{X}_p, X_q - \hat{X}_q) =$$

$$\text{cov}(X_p,X_q) - \text{cov}(X_p,\hat{X}_q) - \text{cov}(\hat{X}_p,X_q) + \text{cov}(\hat{X}_p,\hat{X}_q) =$$

$$\text{cov}(X_p,X_q) - \sum_{X_i \in \mathbf{P_p}} \alpha_{pi}\,\text{cov}(X_p,X_i) - \sum_{X_j \in \mathbf{P_q}} \alpha_{qj}\,\text{cov}(X_q,X_j) + \sum_{X_i \in \mathbf{P_p}} \sum_{X_j \in \mathbf{P_q}} \alpha_{pi}\alpha_{qj}\,\text{cov}(X_i,X_j)$$

Each coefficient of an X_i — X_j edge is a partial correlation, and hence a rational function of Σ. Each of the α_{kj} is a regression coefficient, and hence a rational function of Σ. $\text{cov}(\varepsilon_p,\varepsilon_q)$ is also a polynomial function of Σ, because it is a polynomial function of the

α_{kj} and Σ. Because the parameters of a complete MAG are a polynomial function of Σ, there is a smooth function from Σ to the parameters of a complete MAG.

It was also shown in Theorem 2 that each variable X_k could be written as

$$X_k = \sum_{X_j \in P_k} \alpha_{kj} X_j + \varepsilon_j$$

Hence the function mapping the covariance matrix to the MAG parameters has an inverse, and is one-to-one. In addition, it follows that there is a reduced form for the **X** variables, i.e. they are a rational function of the α_{kj} parameters, the ε variables, and the u-vertices. Hence the covariances among the **X** variables are a rational function of the covariances among the ε variables , the α_{kj} parameters, and the partial correlation among the u-vertices. It follows that there is a smooth function from the parameters of a complete MAG to Σ.

It follows that there is a diffeomorphism from Σ to the parameters of a complete MAG.

There is also a diffeomorphism from the natural parameters of the normal distribution to the covariance matrix of a normal distribution (Kass and Vos, p. 101). The composition of two diffeomophisms is a diffeomorphism (Kass and Vos, p. 101), and hence there is a diffeomorphism from the natural parameters to the parameters of a complete MAG.

Each family of distributions represented by a MAG can be characterized by setting some subset of the parameters of a complete MAG equal to zero. It follows from Theorem 4.2.1 that the distributions represented by a MAG are a curved exponential family.

Since the dimensionality of the full space of k normal variables with zero mean is equal to k(k+1)/2, the dimensionality of a compete MAG is k(k+1)/2. Let M be an incomplete MAG. By Lemma 1, M has a complete extension M', and the dimensionality of M' is k(k+1)/2. Each parameter in M' that is set to zero (one of the α_{kj}, or a covariance between two error terms ε_k and ε_j, or a partial correlation between two u-vertices) corresponds to a pair of variables in M that are not adjacent. The number of parameters in

M is equal to k(k+1)/2 minus the number of parameters in M' set to zero, i.e. k(k+1)/2 minus the number of pairs of variables in M' that are not adjacent to each other. ∴

9) The BIC Score of a Linear MAG

As the sample size increases without limit, the Bayes Information Criterion is an O(1) approximation of a function of the posterior distribution. In the case of a multivariate normal model, for a given sample

$$BIC(M, sample) = -2L(\Sigma_{M(\theta_{max})}, sample) + \ln(samplesize) * df_M,$$

where

- θ_{max} is the maximum likelihood estimate of the parameters for model M from sample,
- $\Sigma_{M(\theta_{max})}$ is the implied covariance matrix for M when θ takes on its maximum likelihood value θ_{max},
- $L(\Sigma_{M(\theta_{max})}, sample)$ is the likelihood of $\Sigma_{M(\theta_{max})}$, and
- df_M is the degrees of freedom (dimensionality) of the MAG M.

(See Raftery, 1993).

10) Example: Noctuid Moth Data

To illustrate the use of these models on a simple data set we present an analysis of data on moth trappings, which originally appeared in the statistical literature in a paper of Cochran (1938), but which were subsequently analyzed by Dempster (1972), who used the data to illustrate covariance selection models, and Whittaker (1990), who fitted a chain graph model to this data. These earlier analyses provide an interesting point of comparison for the partial ancestor graph analysis.

The data consist of one response variable,

moth : log (1 + no. of moths caught in a light trap on one night),

and five covariates:

min : the minimum night temperature,

max : the previous day's maximum temperature,

wind : the average wind speed during the night,

rain : the amount of rain during the night

cloud: the percentage of starlight obscured by clouds

The data as given by Cochran are:

	min	*max*	*wind*	*rain*	*cloud*	*moth*
min	1.00					
max	0.40	1.00				
wind	0.37	0.02	1.00			
rain	0.18	-0.09	0.05	1.00		
cloud	-0.46	0.02	-0.13	-0.47	1.00	
moth	0.29	0.22	-0.24	0.11	-0.37	1.00
Variance	14.03	14.54	2.07	17.11	7.87	3.55

The original observations are not available, but Cochran implies that they come from a complicated design with an effective sample size of 72.

a) Dempster's Model

Dempster (1972) fitted a covariance selection model to this data, which corresponds to the following undirected graph:

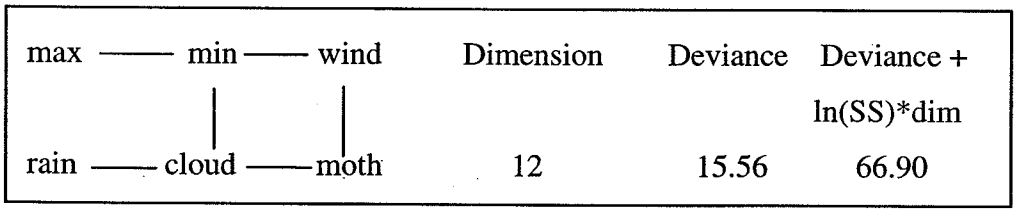


Figure 2: Dempster's Model

where conditional independence is encoded via separation, e.g. *min* $\perp\!\!\!\perp$ *moth* | *cloud,wind.*

Dempster arrived at his model via a forward selection procedure which terminated when it found the first model for which the p-value > 0.05; the p-value was computed by comparing the Deviance to a χ^2 distribution with d.f. = (21 – dimension of the model). We also give Deviance + ln(Sample Size)*Dimension, since this is equal to the BIC score + a constant (note that lower scores correspond to 'better' models under this criterion).

b) *Whittaker's Model*

Whittaker (1990) presents an analysis based on a chain graph, based upon a division of the variables into two blocks, the first containing the five covariates, the second containing the response:

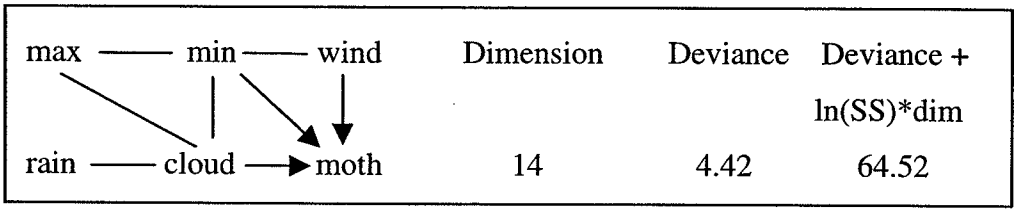


Figure 3: Whittaker's Model[4]

Whittaker arrived at this model by first searching for an undirected model for the covariates, and then regressing *moth* on the five covariates, selecting *min*, *cloud*, and *wind* on the basis of the edge exclusion deviances (which is the deviance of the model with one edge removed against the full model including all covariates). Note that this model implies that *cloud* $\perp\!\!\!\perp$ *wind* | *min*, and does not imply *cloud* $\perp\!\!\!\perp$ *wind* | *min*, *moth* whereas the reverse is true of Dempster's model.

c) *FCI Model*

We applied an algorithm (the FCI algorithm described in Spirtes *et al.* 1993) that searches for sets of MAG models that are statistically equivalent (i.e. represent the same sets of conditional independence relations). Figure 5 shows one of the MAG models represented by the output. (The structural equation modelling programme EQS, developed by Peter Bentler was used to fit these models.)

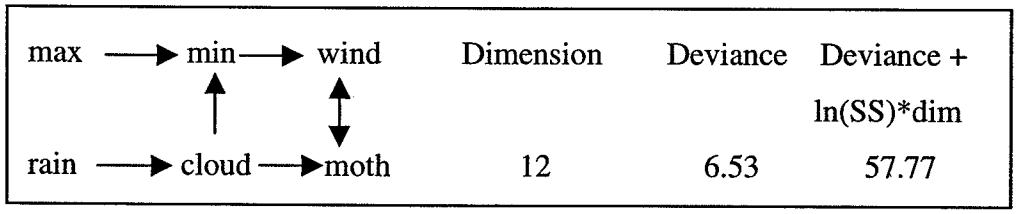


Figure 4: MAG found by FCI search

---

[4]When we used the Bentler's EQS programme to fit this model it gave a deviance of 4.74; Whittaker reports a devianceof 4.42.

This MAG imposes the following conditional independence constraints:

max ⫫ *rain, cloud, moth*;

min ⫫ *rain, moth* | *cloud*;

wind ⫫ *max, cloud, rain* | *min*;

rain ⫫ *max, min, wind, moth* | *cloud.*

This is not a complete list of conditional independences, but it is sufficient to uniquely specify the MAG.

It can be shown that the following structural properties are true of *any* DAG (possibly with latent variables and selection bias) which is conditional independence equivalent to the MAG:

min is not an ancestor of *cloud* or *max*;

wind is not an ancestor of *min* or *moth*;

moth is not an ancestor of *cloud* or *wind*;

min is an ancestor of *wind.*

It is interesting to compare the FCI model to those of Whittaker and Dempster. In fact, the FCI model is nested within Whittaker's model. Since the two models differ by 2 d.f. but the difference in deviance is only 2.11, a likelihood ratio test finds no evidence against the FCI model (p-value 0.348). In fact, the FCI model has the same pairs of adjacent vertices as in Dempster's model. The two extra edges present in Whittaker's model are the *max—cloud* and *min* → *moth* edges. Let us examine these in turn:

In describing how he came up with his model Whittaker states that at first he fitted an undirected model to the covariates, which did not include the *max—cloud* edge, since these two variables are close to being uncorrelated. However, after examining the edge exclusion deviance, which measured the dependence of *max* and *cloud* given *min*, *wind* and *rain* he decided to include this extra edge, since the deviance indicated strong dependence, yet the model without the edge would imply *max* ⫫ *cloud* | *min, wind, rain.* The FCI model manages to accomodate both the marginal independence and the conditional independence. In fact, in this case a DAG model such as shown in Figure 6 could also have achieved this.

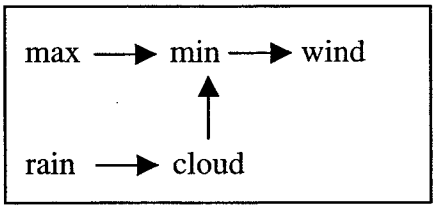


Figure 5: DAG model for the covariates

This calls into question the motivation for blocking variables and fitting undirected graphs within blocks, and directed edges between blocks, that is advocated by Whittaker and others.

If we now examine the *min* → *moth* edge that is absent in the FCI MAG, but present in Whittaker's model, this illustrates a potential shortcoming of regressing a response on all previous covariates in order to determine those that are causes of the response. Consider the DAG with latent variables T_1, T_2, shown in Figure 8. This DAG is conditional independence equivalent to the FCI MAG over the variables {*cloud*, *min*, *wind*, *moth*}. Further, it is compatible with the background knowledge that Whittaker used when constructing his model: all the covariates temporally preceed *moth*. However, although *min* and *moth* are not directly related in this DAG, *min* and *moth* are dependent given the other covariates *cloud* and *wind*.

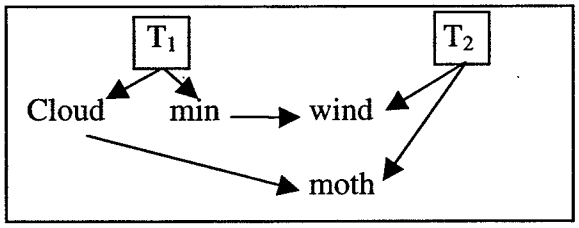


Figure 6: A DAG with latent variables

It is well known that failing to include a confounding variable in a regression may lead to a spurious dependence between two variables. What is perhaps less well known is that *including* the wrong variable in a regression may lead to a spurious dependence: in this case regressing *moth* on *min*, *wind* and *cloud* leads to a spurious dependence between *moth* and *min*, and thus to the additional edge in Whittaker's model.

It should be stressed that in comparing the FCI model to Whittaker's model we do not wish to imply that the FCI model is the 'true' model for this dataset. With a

comparatively small sample size, as in this case, we would not expect the data to uniquely identify a single model: this is borne out by the fact that there are many different PAG models with scores that are relatively close. (See Figure 10.) The existence of so many different models with relatively similar scores must temper any causal or structural inferences that we might wish to draw from this analysis, unless all of the models receiving high scores share this feature in common.

The FCI search is a heuristic search procedure based upon the results of a series of conditional independence tests, and is not guaranteed to find the (set of) MAGs with the best BIC score (though it will do so asymptotically). However, it appears that in this example the FCI algorithm did locate the MAG with the best score; a greedy search failed to find a MAG with a higher score. A number of other MAGs, together with the associated deviance and scores are given in Figure 7.

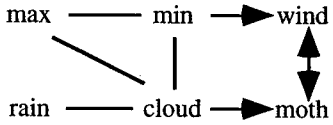
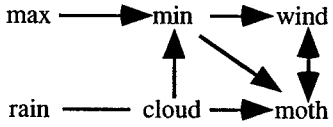
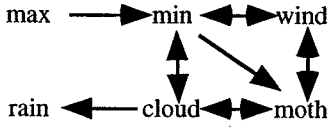
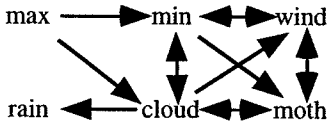
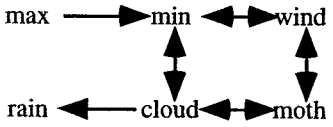
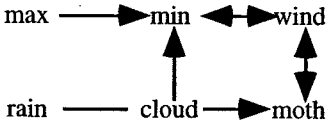
	Dimension	Deviance	Deviance +ln(SS)*dim.
	13	6.50	62.01
	13	4.77	60.28
	13	4.77	59.11
	15	2.26	66.31
	12	11.13	62.37
	12	7.52	58.76

Figure 7: Other MAG Models

11) References

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

Cochran, W.G. (1938). The omission or addition of an independent variate in multiple linear regression. JRSS Supplement, 5, pp.171-176.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models in Proc. Sixth Conference on Uncertainty in AI. Association for Uncertainty in AI, Inc., Mountain View, CA.

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

Wright, S. (1934). The method of path coefficients. Annals of Mathematical Statistics, 5, 161-215.