Scoring Ancestral Graph Models

Thomas Richardson & Peter Spirtes

January 19, 1999

Technical Report No. CMU-PHIL-98

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Scoring Ancestral Graph Models

Thomas Richardson, University of Washington (tsr@stat.washington.edu)
Peter Spirtes, Carnegie Mellon University.

I. Introduction

There has recently been significant progress in the development of algorithms for learning the directed acyclic graph (DAG) part of a Bayesian network without latent variables from data and optional background knowledge. However, the problem of learning the DAG part of a Bayesian network with latent (unmeasured) variables is much more difficult for two reasons: first the number of possible models is infinite, and second, calculating scores for latent variables models is generally much slower than calculating scores for models without latent variables.

In this paper we will describe how to extend search algorithms developed for non-latent variable DAG models to the case of DAG models with latent variables. We will introduce two generalizations of DAGs, called mixed ancestor graphs (or MAGs) and partial ancestor graphs (or PAGs), and briefly describe how they can be used to search for latent variable DAG models. In the last section we apply these techniques to a dataset concerning noctuid moth trappings, that was previoulsy analyzed using models based on undirected graphs and chain graphs.

II. Directed Acyclic Graphs (DAGs)

A Bayesian network consists of two distinct parts: a directed acyclic graph (DAG or belief-network structure) and a set of parameters for the DAG. Under the statistical interpretation of a DAG, a DAG with a set of vertices **V** represents a set of probability measures over **V**. (We place sets of variables and defined terms in boldface.) Following the terminology of Lauritzen *et al.* (1990) say that a probability measure over a set of variables **V** satisfies the **local directed Markov property** for a directed acyclic graph (or DAG) G with vertices **V** if and only if for every W in **V**, W is independent of **V**\(**Descendants**(W) $\cup$ **Parents**(W)) given **Parents**(W), where **Parents**(W) is the set of parents of W in G, and **Descendants**(W) is the set of descendants of W in **G**. (Note that a vertex is its own ancestor and descendant, although not its own parent or child.) A DAG G **represents** the set of probability measures which satisfy the local directed Markov property for G. Variants of probabilistic DAG models were introduced in the 1980's in Pearl (1988) among others. Many familiar parametric models, such as recursive structural equation models with uncorrelated errors, factor analytic models, item response models, etc. are special cases of parameterized DAGs. (See Pearl 1988 for references.)

Under the causal interpretation, a DAG represents the causal relations in a given population with a set of vertices **V** when there is an edge from A to B if and only if A is a direct cause of B relative to **V**. The use of DAGs to simultaneously represent a set of causal hypotheses and a family of probability distributions extends back to the path diagrams introduced by Sewell Wright (1934). For the class of models considered in this paper we make two assumptions relating causal DAGs to probability distributions.

Causal Independence Assumption: If A does not cause B, and B does not cause A, and there is no third variable that causes both A and B, then A and B are independent.

Causal Faithfulness Assumption: If a causal DAG *M* correctly describes the causal structure in a population with probability distribution P, then each conditional independence true in P is entailed by *M*.

These assumptions linking the statistical and causal interpretations of DAGs are defended in Spirtes, Glymour and Scheines (1993).

III. Partial Ancestral Graphs (PAGs)

In some cases, not all of the variables in a DAG can be measured. We call those variables whose values are measured the **observed** variables, and all other variables in the DAG **latent** variables. For a given division of the variables in a DAG G into observed and latent, we write $G(\mathbf{O},\mathbf{L})$ where **O** is the set of observed variables and **L** is the set of latent variables.

A DAG *G* **entails a conditional independence relation** if and only if it is true in every probability measure satisfying the local directed Markov property for *G*. Two directed graphs $G_1(\mathbf{O},\mathbf{L})$ and $G_2(\mathbf{O'},\mathbf{L'})$ are **conditional independence equivalent** if and only if $\mathbf{O} = \mathbf{O'}$, and for all **X**, **Y** and **Z** included in **O**, $G_1(\mathbf{O},\mathbf{L})$ entails **X** and **Y** are independent conditional on **Z** if and only if $G_2(\mathbf{O},\mathbf{L})$ entails **X** and **Y** are independent conditional on **Z**. We denote the set of directed acyclic graphs that are conditional independence equivalent to $G(\mathbf{O},\mathbf{L})$ as **Equiv**$(G(\mathbf{O},\mathbf{L}))$.

A **partial ancestral graph** (PAG) can be used to represent any subset of **Equiv**$(G(\mathbf{O},\mathbf{L}))$. A PAG is an extended graph consisting of a set of vertices **O**, and a set of edges between vertices, where there may be the following kinds of edges: A $\leftrightarrow$ B, A o—o B, A o$\rightarrow$ B, A $\leftarrow$o B, A $\rightarrow$ B or A $\leftarrow$ B. We say that the A endpoint of A $\rightarrow$ B is "–"; the A endpoint of an A $\leftrightarrow$ B, A $\leftarrow$o B, or A $\leftarrow$ B edge is "<"; and the A endpoint of a A o—o B or A o$\rightarrow$ B is "o". The conventions for the B endpoints are analogous. In addition pairs of edge endpoints may be connected by underlining (interpreted below). A partial ancestral graph for a set of directed acyclic graphs **G** each sharing the same set of observed variables **O**, contains partial information about the ancestor relations in **G**, namely only those ancestor relations common to all members of **G**. (If we allow **G** to contain directed cyclic graphs as well as directed acyclic graphs then several extra types of

edges are needed in the PAG (See Richardson, 1996). In the following definition, which provides a semantics for PAGs we use "*" as a meta-symbol indicating the presence of any one of {o, –, >}, e.g. A *→ B represents either A → B, A ↔ B, or A o→ B.

Partial Ancestral Graphs (PAGs)

If **G** is a set of directed acyclic graphs included in **Equiv**(G(**O**, **L**)), Ψ (with vertices **O**) is a PAG for **G** if and only if

(i) There is an edge between A and B in Ψ if and only if every DAG in **G** does not entail that A and B are independent conditional on any subset of **O**\{A,B}.

(ii) If there is an edge in Ψ out of A, i.e. A → B, then A is an ancestor of B in every graph in **G**.

(iii) If there is an edge in Ψ into B, i.e. A*→ B, then in every DAG in **G**, B is **not** an ancestor of A.

(iv) If there is an underlining A*—*B*—*C in Ψ then B is an ancestor of (at least one of) A or C in every DAG in **G**.

(v) Any edge endpoint not marked in one of the above ways is left with a small circle thus: o—*.

Some examples of PAGs are shown in Figure 1, where **O** = {A,B,C,D}. In cases where the distinction between latent variables and measured variables is important, we enclose latent variables in ovals. (The MAGs in Figure 1 are defined in the next section.)

The requirement that **G** is included in **Equiv**(G(**O**, **L**)) guarantees that if one directed acyclic graph in **Equiv**(G(**O**, **L**)) does not entail that A and B are independent conditional on any subset of **O**\{A,B}, then all directed acyclic graphs in **Equiv**(G(**O**, **L**)) do not entail that A and B are independent conditional on any subset of **O**\{A,B}.

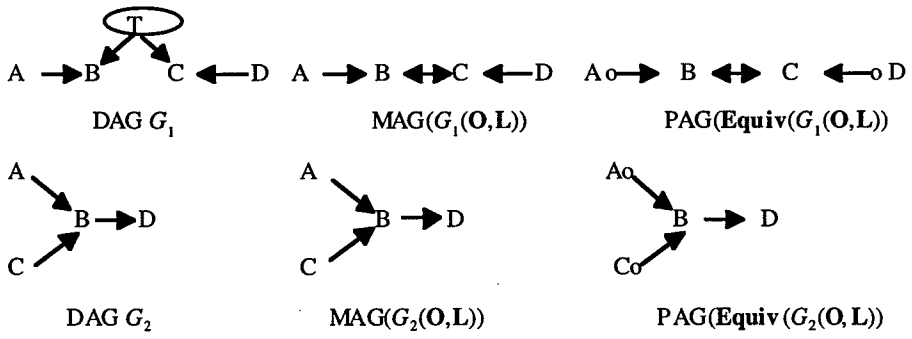


Figure 1

Note that only condition (i) gives necessary and sufficient conditions about features of the PAG. All of the other conditions are merely necessary conditions. That means that there can be more than one PAG representing a given set **G**; two such PAGs have the same

adjacencies, but one may contain a "o" endpoint where the other contains a "–" or "> " endpoint. There are PAGs for **Equiv**(G(**O**, **L**)) with enough orientation information to determine whether or not each DAG in **Equiv**(G(**O**, **L**)) entails that **A** and **B** are independent conditional on any subset included in **O**\(**A** ∪ **B**); we will say that any such PAG that has enough orientations to do this is "weakly complete" for **Equiv**(G(**O**, **L**)). (Weak completeness does *not* entail that every ancestor relation common to every member of **Equiv**(G(**O**, **L**)) is explicitly represented in the PAG.)

Thus a PAG can be used to represent both the ancestor relations among the members of **O** common to members of **G**, and the set of conditional independence relations among the members of **O** in **G**. Some PAGs (e.g. PAG(**Equiv**(G_1(**O**,**L**))) in Figure 1) represent a set of conditional independence relations not entailed by any DAG G(**O**,**L**) where **L** = ∅.

PAGs have two distinct uses. Just as DAGs can be used by algorithms to perform fast conditionalizations, PAGs can be used in a similar way. And just as, given a causal interpretation, DAGs can be used to calculate the effects of any ideal intervention upon a system, PAGs, given a causal interpretation, can be used to calculate the effects of *some* ideal interventions upon a system. (See Spirtes et al. 1993, where PAGs are called POIPGs.)

While it would generally be preferable to know the true causal DAG G(**O**,**L**) rather than a PAG representing **Equiv**(G(**O**, **L**)), there are several reasons why it may be easier to find a PAG representing **Equiv**(G(**O**, **L**)) than it is to find G(**O**,**L**) itself. First the space of PAGs is finite, while the space of DAGs with latent variables is infinite. Second, for a variety of scores for models (such as BIC, posterior probability, etc.) there may be many different DAGs which receive the same score, but represent different causal theories and make different predictions about the effects of interventions upon a system. The data alone does not allow one to distinguish between these models, so even with population data, one cannot be sure which is the correct causal model. Nevertheless, for some (but not all) equivalence classes of causal models, and some (but not all) ideal interventions, it is possible to use a PAG to consistenly estimate the effect of the intervention, even without knowing which causal model represented by the PAG is the correct model. Note that this strategy is not useful in instances where every pair of measured variables has some strong latent common cause; in that case the PAG that represents **Equiv**(G(**O**, **L**)) is completely connected, and cannot be used to predict the effects of any ideal interventions on the system.

Is it possible to find a PAG from data and background knowledge? The FCI algorithm, under a set of assumptions described in Spirtes et al. 1993, is guaranteed in the large sample limit to find a weakly complete correct PAG for a given distribution. It uses a series of conditional independence tests to construct a PAG that represents a given distribution. The algorithm is exponential in the number of vertices in the PAG in the worst case (as is any algorithm based upon conditional independence tests.) However, the large

sample reliability does not guarantee reliability on realistic sample sizes, and if the power of the conditional independence tests is low, the results of the tests are not compatible with any single PAG. For these reasons, it would be desirable to have a search that was not based upon conditional independence tests, or could be used to supplement an algorithm based upon conditional independence tests by using the output of the FCI algorithm as a starting point for a search.

Recently, a number of algorithms for searching for DAGs without latent variables have been developed that do not rely on conditional independence tests. (Chickering et al. 1995, Spirtes and Meek 1995) Instead, these are heuristic searches that attempt to maximize a score. We will describe here a heuristic PAG search that attempts to find a PAG with the highest score. One problem with this approach is that because a PAG represents a set of DAG models which may receive different scores (either Bayes Information Criterion, posterior probability, etc.) a PAG cannot be assigned a score by setting its score equal to an arbitrarily chosen DAG that it represents. In the next section we will show how to indirectly assign a score to a PAG.

IV. Mixed Ancestral Graphs (MAGs)

A MAG (or mixed ancestral graph) is a completely oriented PAG for a set of graphs which consists of a single directed acyclic graph $G(\mathbf{O},\mathbf{L})$. (By completely oriented we mean that there are no "o" endpoints on any edge). Some examples of MAGs are shown in Figure 1, where $\mathbf{O} = \{A,B,C,D\}$.

A MAG can also be considered a representation of a set of conditional independence relations among variables in $\mathbf{O}$ (which in some cases cannot be represented by any DAG containing just variables in $\mathbf{O}$; e.g. MAG($G_1(\mathbf{O},\mathbf{L})$) in Figure 1.) A MAG imposes no restrictions on the set of distributions it represents other than the conditional independence relations that it entails. (The class of MAGs is neither a subset nor a superset of other generalizations of DAGs such as chain graphs, cyclic directed graphs, or cyclic chain graphs.)

MAGs have the following useful features:

- DAG G_1 in Figure 1 is an example of a DAG such that as the sample size increases without limit, the difference between the Bayes Information Criterion (BIC) of MAG(G_1,$\mathbf{O}$) and the BIC of any DAG G' that contains only variables in $\mathbf{O}$ increases without limit almost surely. Hence in some cases a maximum likelihood estimate of the MAG parameters is a better estimator of some of the population parameters than the maximum likelihood estimate of any DAG parameters.

- In the large sample limit, for multi-variate normal or discrete distributions, any (possibly latent variable) DAG with a maximum BIC score is represented by the MAG with the highest BIC score among all MAGs.
- There is a three place graphical relation among disjoint sets of vertices (**A** is d-separated from B given **C**) which holds if and only if the MAG entails that **A** is independent of **B** conditional on **C**. D-separation in MAGs is a simple extension of Pearl's d-separation relation (Pearl 1988) defined over DAGs.

If a PAG Ψ represents **Equiv**(G(**O**,**L**)), we say that any MAG that represents graph G(**O**,**L**) is represented by Ψ. For every PAG, there is some MAG that it represents, and every MAG represented by a PAG receives the same BIC score. Thus a PAG can be assigned a score by finding some MAG that it represents, scoring the MAG, and assigning that score to the PAG. It is possible that a PAG represents some non-MAG model that receives a higher BIC score than any MAG represented by the PAG. However, assigning a MAG score to a PAG that represents it has the following desirable property. For any distribution P(**O**), if there is some DAG G that contains **O**, such that for any three disjoint sets of variables **X**, **Y**, **Z** $\subseteq$ **O**, **X** is independent of **Y** given **Z** if and only if **X** is d-separated from **Y** given **Z** in G, then **P**(**O**) is said to be **faithful** to G over **O**. For any multi-variate normal distribution P(**O**), if P(**O**) is faithful to some DAG G over **O**, then in the large sample limit the PAG that represents G receives the highest BIC score among all PAGs.

A. *Parameterizing Gaussian MAGs*

We will describe how a parameterization of a MAG in the multi-variate normal case is an extension of a parameterization of a DAG corresponding to a "structural equation model". (Parameterization and estimation of parameters in the case of discrete variables is an area of current research.)

The variables in a linear structural equation model (SEM) can be divided into two sets, the "error variables" or "error terms," and the substantive variables. Corresponding to each substantive variable X_i is a linear equation with X_i on the left hand side of the equation, and the direct causes of X_i plus the error term ε_i on the right hand side of the equation. Since we have no interest in first moments, without loss of generality each variable can be expressed as a deviation from its mean.

Consider, for example, two SEMs S_1 and S_2 over **X** = $\{X_1, X_2, X_3\}$, where in both SEMs X_1 is a direct cause of X_2. The structural equations in Figure 2 are common to both S_1 and S_2:

$$X_1 = \varepsilon_1$$

$$X_2 = \beta_{21} X_1 + \varepsilon_2$$

$$X_3 = \varepsilon_3$$

Figure 2: Structural Equations for SEMs S_1 and S_2

where β_{21} is a free parameters ranging over real values, and ε_1, ε_2 and ε_3 are error terms. In addition suppose that ε_1, ε_2 and ε_3 are distributed as multivariate normal. In S_1 we will assume that the correlation between each pair of distinct error terms is fixed at zero. The free parameters of S_1 are $\theta = <\beta, \mathbf{P}>$, where β is the set of linear coefficients $\{\beta_{21}\}$ and **P** is the set of variances of the error terms. We will use $\Sigma_{S_1(\theta_1)}$ to denote the covariance matrix parameterized by the vector θ_1 for model S_1. If all the pairs of error terms in a SEM S are uncorrelated, we say S is a SEM with **uncorrelated errors**.

S_2 contains the same structural equations as S_1, but in S_2 we will allow the errors between X_2 and X_3 to be correlated, i.e., we make the correlation between the errors of X_2 and X_3 a free parameter, instead of fixing it at zero, as in S_1. In S_2 the free parameters are $\theta = <\beta, \mathbf{P'}>$, where β is the set of linear coefficients $\{\beta_{21}\}$ and **P'** is the set of variances of the error terms and the correlation between ε_2 and ε_3. If the correlations between any of the error terms in a SEM are not fixed at zero, we will call it a SEM with **correlated errors**.

It is possible to associate with each SEM with uncorrelated errors a directed graph that represents the causal structure of the model and the form of the linear equations. For example, the directed graph associated with the substantive variables in S_1 is $X_1 \rightarrow X_2$ X_3, because X_1 is the only substantive variable that occurs on the right hand side of the equation for X_2.

It is generally accepted that correlation is to be explained by some form of causal connection. Accordingly if ε_2 and ε_3 are correlated we will assume that either ε_2 causes ε_3, ε_3 causes ε_2, some latent variable causes both ε_2 and ε_3, or some combination of these. We represent the correlated error between ε_2 and ε_3 by introducing a latent variable T that is a common cause of X_2 and X_3. If $\mathbf{O} = \{X_1, X_2, X_3\}$, the MAG for the directed graph associated with S_2 is $X_1 \rightarrow X_2 \leftrightarrow X_3$. The statistical justification for this is provided in Spirtes et al. (1996). It turns out that the set of MAGs is a subset of the set of recursive structural equation models with correlated errors. Hence, there are well known techniques (Bollen, 1992) for estimating and performing statistical tests upon MAG models such as S_2.

B. *The Bayes Information Criterion (BIC) Score of a 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 multi-variate normal structural equation model, for a given sample

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

$$= \text{constant} + \text{Deviance(M)} + \ln(\text{samplesize}) * df_M$$

where

- θ_{max} is the maximum likelihood estimate of the parameters for model M from sample,
- $\Sigma_{M(\theta_{max})}$ is the covariance matrix for M when θ takes on its maximum likelihood value θ_{max},
- $L(\Sigma_{M(\theta_{max})}, \text{sample})$ is the likelihood ratio test statistic of $\Sigma_{M(\theta_{max})}$,
- df_M is the degrees of freedom of the MAG M.
- Deviance(M) is twice the difference between the unconstrained maximum of the log-likelihood and the maximum taken over M.

(See Raftery, 1993).

V. 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

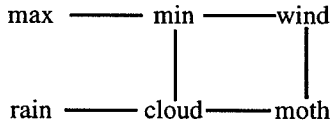
The data as given by Cochran are:

	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:



Dimension	Deviance	Deviance +ln(SS)*dim.
12	15.66	66.90

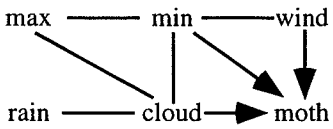
Figure 3: Dempster's Model

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 was greater than 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:



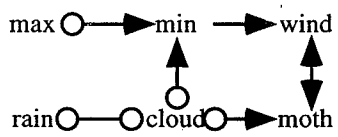
Dimension	Deviance	Deviance +ln(SS)*dim.
14	4.42	64.52

Figure 4: Whittaker's Model[1]

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* ⫫ *wind* | *min*, and does not imply *cloud* ⫫ *wind* | *min*, *moth* whereas the reverse is true of Dempster's model.

C. *FCI Model*

Applying the FCI algorithm to this data resulted in the Partial Ancestral Graph shown in Figure 5. (The structural equation modelling programme EQS, developed by Peter Bentler, was used to fit these models.)



Dimension	Deviance	Deviance +ln(SS)*dim.
12	6.53	57.77

Figure 5: PAG found by FCI search

This PAG 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 PAG. Further, the PAG implies the following structural properties are true of any DAG (possibly with latent variables) which is conditional independence equivalent to the PAG:

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*;

---

[1] When we used the Bentler's EQS programme to fit this model we obtained a deviance of 4.17 while Whittaker reports a devianceof 4.42.

min is an ancestor of *wind*;

there is an unmeasured common cause of *moth* and *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 = 0.39). 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 $max \perp\!\!\!\perp cloud \mid min, wind, rain$. The FCI model manages to accommodate 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.

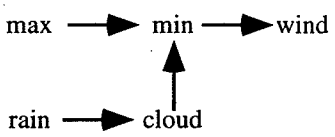


Figure 6: DAG model for the covariates

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 PAG, 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 PAG 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 precede *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*.

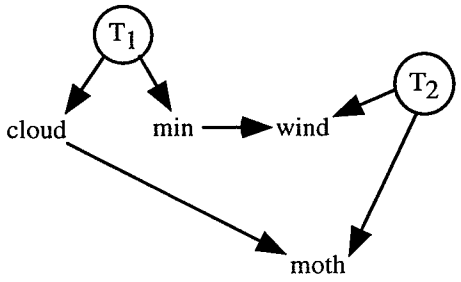


Figure 8: A DAG with latent variables

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.

As described in section IV, it is not possible to score a PAG directly, but only through parameterizing a MAG represented by it. The MAG used for scoring the FCI PAG is shown in Figure 9.

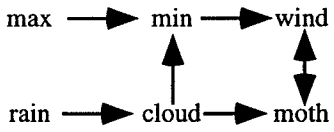


Figure 9: MAG corresponding to PAG in Fig. 5

Note that the same BIC score would be obtained from scoring any other MAG represented by the PAG. 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 PAG with the best BIC score (though it will do so asymptotically). However, it appears that in this example the FCI algorithm did locate the PAG with the best score; a greedy search failed to find a PAG with a higher score. A number of other PAGs, together with the associated deviance and scores are given in Figure 10.

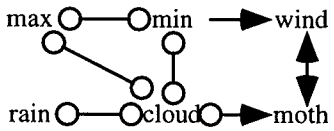
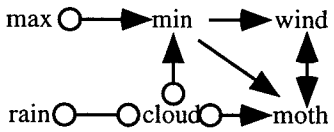
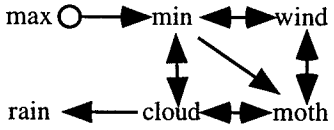
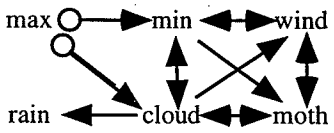
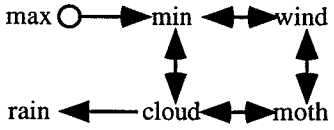
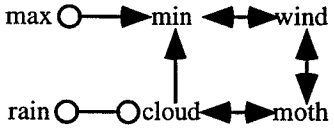
	Dimension	Deviance	Deviance +ln(SS)*dim.
	13	6.50	62.01
	13	4.77	60.28
	13	4.77	59.11
	15	2.26	66.31
	12	11.13	62.37
	12	7.52	58.76

Figure 9: Other PAG Models

VI. *Summary and Future Work*

In this paper have introduced Partial Ancestral Graphs as a representation for equivalence classes of DAGs with latent variables, that captures structural features that are common to all the DAGs in a given equivalence class. We have presented a method for parameterizing the set of Gaussian distributions defined by these conditional independence constraints. This allows a BIC score to be calculated for a PAG. Finally, we have illustrated through an example, that the class of PAG models allow greater flexibility in representing conditional independence, leading to more parsimonious models. Research is currently focused upon

parameterizing MAGs with discrete variables, and developing fast algorithms for transforming PAGs into MAGs.

VII. *Bibliography*

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.

Chickering, D. and Geiger, D. and Heckerman, D. (1995). Learning Bayesian networks: Search methods and experimental results. Preliminary papers of the fifth international workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL, pp. 112-128.

Dempster, A.P. (1972). Covariance Selection. Biometrics, 28, pp. 157-175.

Lauritzen, S., Dawid, A., Larsen, B., Leimer, H., (1990) Independence properties of directed Markov fields, Networks, 20, 491-505.

Pearl, J., (1988). Probabilistic Reasoning in Intelligent Systems, Morgan Kaufman: San Mateo, CA.

Pearl, J. (1995) Causal diagrams for empirical research, Biometrika, 82.

Raftery, A. (1993) Bayesian Model Selection in Structural Equation Models, in Testing Structural Equation Models, ed. by K. Bollen and S. Long, Sage Publications.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. Uncertainty in Artificial Intelligence, Proceedings, 12th Conference, Morgan Kaufman, CA.

Spirtes, P., Glymour, C., and Scheines, R., (1993) Causation, Prediction, and Search, (Springer-Verlag Lecture Notes in Statistics 81, New York).

Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1996). Using D-separation to Calculate Zero Partial Correlations in Linear Models with Correlated Errors, Carnegie Mellon University Technical Report Phil-72.

Spirtes, P., and Meek, C. (1995). Learning Bayesian Networks with Discrete Variables from Data", in Proceedings of The First International Conference on Knowledge Discovery and Data Mining, ed. by Usama M. Fayyad and Ramasamy Uthurusamy, AAI Press, pp. 294-299.

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.

Scoring Ancestral Graph Models

Thomas Richardson, University of Washington (tsr@stat.washington.edu)

Peter Spirtes, Carnegie Mellon University.

I. Introduction

There has recently been significant progress in the development of algorithms for learning the directed acyclic graph (DAG) part of a Bayesian network without latent variables from data and optional background knowledge. However, the problem of learning the DAG part of a Bayesian network with latent (unmeasured) variables is much more difficult for two reasons: first the number of possible models is infinite, and second, calculating scores for latent variables models is generally much slower than calculating scores for models without latent variables.

In this paper we will describe how to extend search algorithms developed for non-latent variable DAG models to the case of DAG models with latent variables. We will introduce two generalizations of DAGs, called mixed ancestor graphs (or MAGs) and partial ancestor graphs (or PAGs), and briefly describe how they can be used to search for latent variable DAG models. In the last section we apply these techniques to a dataset concerning noctuid moth trappings, that was previoulsy analyzed using models based on undirected graphs and chain graphs.

II. Directed Acyclic Graphs (DAGs)

A Bayesian network consists of two distinct parts: a directed acyclic graph (DAG or belief-network structure) and a set of parameters for the DAG. Under the statistical interpretation of a DAG, a DAG with a set of vertices **V** represents a set of probability measures over **V**. (We place sets of variables and defined terms in boldface.) Following the terminology of Lauritzen *et al.* (1990) say that a probability measure over a set of variables **V** satisfies the **local directed Markov property** for a directed acyclic graph (or DAG) G with vertices **V** if and only if for every W in **V**, W is independent of **V**\(**Descendants**(W) $\cup$ **Parents**(W)) given **Parents**(W), where **Parents**(W) is the set of parents of W in G, and **Descendants**(W) is the set of descendants of W in **G**. (Note that a vertex is its own ancestor and descendant, although not its own parent or child.) A DAG G **represents** the set of probability measures which satisfy the local directed Markov property for G. Variants of probabilistic DAG models were introduced in the 1980's in Pearl (1988) among others. Many familiar parametric models, such as recursive structural equation models with uncorrelated errors, factor analytic models, item response models, etc. are special cases of parameterized DAGs. (See Pearl 1988 for references.)

Under the causal interpretation, a DAG represents the causal relations in a given population with a set of vertices **V** when there is an edge from A to B if and only if A is a direct cause of B relative to **V**. The use of DAGs to simultaneously represent a set of causal hypotheses and a family of probability distributions extends back to the path diagrams introduced by Sewell Wright (1934). For the class of models considered in this paper we make two assumptions relating causal DAGs to probability distributions.

Causal Independence Assumption: If A does not cause B, and B does not cause A, and there is no third variable that causes both A and B, then A and B are independent.

Causal Faithfulness Assumption: If a causal DAG *M* correctly describes the causal structure in a population with probability distribution P, then each conditional independence true in P is entailed by *M*.

These assumptions linking the statistical and causal interpretations of DAGs are defended in Spirtes, Glymour and Scheines (1993).

III. Partial Ancestral Graphs (PAGs)

In some cases, not all of the variables in a DAG can be measured. We call those variables whose values are measured the **observed** variables, and all other variables in the DAG **latent** variables. For a given division of the variables in a DAG G into observed and latent, we write $G(\mathbf{O},\mathbf{L})$ where **O** is the set of observed variables and **L** is the set of latent variables.

A DAG *G* **entails a conditional independence relation** if and only if it is true in every probability measure satisfying the local directed Markov property for *G*. Two directed graphs $G_1(\mathbf{O},\mathbf{L})$ and $G_2(\mathbf{O'},\mathbf{L'})$ are **conditional independence equivalent** if and only if $\mathbf{O} = \mathbf{O'}$, and for all **X**, **Y** and **Z** included in **O**, $G_1(\mathbf{O},\mathbf{L})$ entails **X** and **Y** are independent conditional on **Z** if and only if $G_2(\mathbf{O},\mathbf{L})$ entails **X** and **Y** are independent conditional on **Z**. We denote the set of directed acyclic graphs that are conditional independence equivalent to $G(\mathbf{O},\mathbf{L})$ as $\mathbf{Equiv}(G(\mathbf{O},\mathbf{L}))$.

A **partial ancestral graph** (PAG) can be used to represent any subset of $\mathbf{Equiv}(G(\mathbf{O},\mathbf{L}))$. A PAG is an extended graph consisting of a set of vertices **O**, and a set of edges between vertices, where there may be the following kinds of edges: A $\leftrightarrow$ B, A o—o B, A o$\rightarrow$ B, A $\leftarrow$o B, A $\rightarrow$ B or A $\leftarrow$ B. We say that the A endpoint of A $\rightarrow$ B is "–"; the A endpoint of an A $\leftrightarrow$ B, A $\leftarrow$o B, or A $\leftarrow$ B edge is "<"; and the A endpoint of a A o—o B or A o$\rightarrow$ B is "o". The conventions for the B endpoints are analogous. In addition pairs of edge endpoints may be connected by underlining (interpreted below). A partial ancestral graph for a set of directed acyclic graphs **G** each sharing the same set of observed variables **O**, contains partial information about the ancestor relations in **G**, namely only those ancestor relations common to all members of **G**. (If we allow **G** to contain directed cyclic graphs as well as directed acyclic graphs then several extra types of

edges are needed in the PAG (See Richardson, 1996). In the following definition, which provides a semantics for PAGs we use "*" as a meta-symbol indicating the presence of any one of {o, –, >}, e.g. A *→ B represents either A → B, A ↔ B, or A o→ B.

Partial Ancestral Graphs (PAGs)

If **G** is a set of directed acyclic graphs included in **Equiv**(G(**O**, **L**)), Ψ (with vertices **O**) is a PAG for **G** if and only if

(i) There is an edge between A and B in Ψ if and only if every DAG in **G** does not entail that A and B are independent conditional on any subset of **O**\{A,B}.

(ii) If there is an edge in Ψ out of A, i.e. A → B, then A is an ancestor of B in every graph in **G**.

(iii) If there is an edge in Ψ into B, i.e. A*→ B, then in every DAG in **G**, B is **not** an ancestor of A.

(iv) If there is an underlining A*—<u>*B*</u>—*C in Ψ then B is an ancestor of (at least one of) A or C in every DAG in **G**.

(v) Any edge endpoint not marked in one of the above ways is left with a small circle thus: o—*.

Some examples of PAGs are shown in Figure 1, where **O** = {A,B,C,D}. In cases where the distinction between latent variables and measured variables is important, we enclose latent variables in ovals. (The MAGs in Figure 1 are defined in the next section.)

The requirement that **G** is included in **Equiv**(G(**O**, **L**)) guarantees that if one directed acyclic graph in **Equiv**(G(**O**, **L**)) does not entail that A and B are independent conditional on any subset of **O**\{A,B}, then all directed acyclic graphs in **Equiv**(G(**O**, **L**)) do not entail that A and B are independent conditional on any subset of **O**\{A,B}.

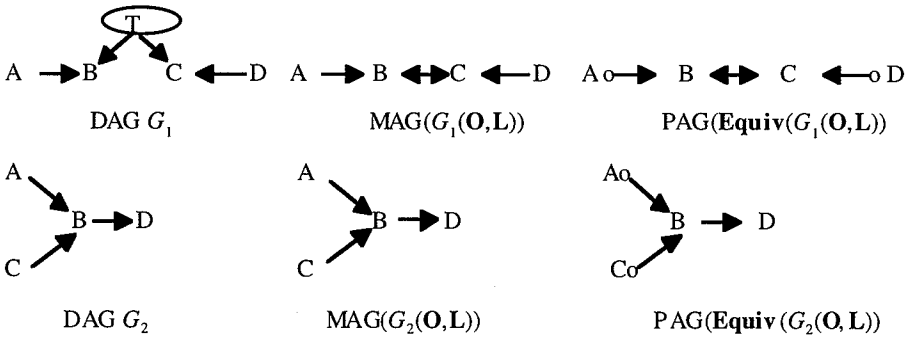


Figure 1

Note that only condition (i) gives necessary and sufficient conditions about features of the PAG. All of the other conditions are merely necessary conditions. That means that there can be more than one PAG representing a given set **G**; two such PAGs have the same

adjacencies, but one may contain a "o" endpoint where the other contains a "–" or "> " endpoint. There are PAGs for **Equiv**(G(**O**, **L**)) with enough orientation information to determine whether or not each DAG in **Equiv**(G(**O**, **L**)) entails that **A** and **B** are independent conditional on any subset included in **O**\(**A** ∪ **B**); we will say that any such PAG that has enough orientations to do this is "weakly complete" for **Equiv**(G(**O**, **L**)). (Weak completeness does *not* entail that every ancestor relation common to every member of **Equiv**(G(**O**, **L**)) is explicitly represented in the PAG.)

Thus a PAG can be used to represent both the ancestor relations among the members of **O** common to members of **G**, and the set of conditional independence relations among the members of **O** in **G**. Some PAGs (e.g. PAG(**Equiv**(G_1(**O**,**L**))) in Figure 1) represent a set of conditional independence relations not entailed by any DAG G(**O**,**L**) where **L** = ∅.

PAGs have two distinct uses. Just as DAGs can be used by algorithms to perform fast conditionalizations, PAGs can be used in a similar way. And just as, given a causal interpretation, DAGs can be used to calculate the effects of any ideal intervention upon a system, PAGs, given a causal interpretation, can be used to calculate the effects of *some* ideal interventions upon a system. (See Spirtes et al. 1993, where PAGs are called POIPGs.)

While it would generally be preferable to know the true causal DAG G(**O**,**L**) rather than a PAG representing **Equiv**(G(**O**, **L**)), there are several reasons why it may be easier to find a PAG representing **Equiv**(G(**O**, **L**)) than it is to find G(**O**,**L**) itself. First the space of PAGs is finite, while the space of DAGs with latent variables is infinite. Second, for a variety of scores for models (such as BIC, posterior probability, etc.) there may be many different DAGs which receive the same score, but represent different causal theories and make different predictions about the effects of interventions upon a system. The data alone does not allow one to distinguish between these models, so even with population data, one cannot be sure which is the correct causal model. Nevertheless, for some (but not all) equivalence classes of causal models, and some (but not all) ideal interventions, it is possible to use a PAG to consistenly estimate the effect of the intervention, even without knowing which causal model represented by the PAG is the correct model. Note that this strategy is not useful in instances where every pair of measured variables has some strong latent common cause; in that case the PAG that represents **Equiv**(G(**O**, **L**)) is completely connected, and cannot be used to predict the effects of any ideal interventions on the system.

Is it possible to find a PAG from data and background knowledge? The FCI algorithm, under a set of assumptions described in Spirtes et al. 1993, is guaranteed in the large sample limit to find a weakly complete correct PAG for a given distribution. It uses a series of conditional independence tests to construct a PAG that represents a given distribution. The algorithm is exponential in the number of vertices in the PAG in the worst case (as is any algorithm based upon conditional independence tests.) However, the large

sample reliability does not guarantee reliability on realistic sample sizes, and if the power of the conditional independence tests is low, the results of the tests are not compatible with any single PAG. For these reasons, it would be desirable to have a search that was not based upon conditional independence tests, or could be used to supplement an algorithm based upon conditional independence tests by using the output of the FCI algorithm as a starting point for a search.

Recently, a number of algorithms for searching for DAGs without latent variables have been developed that do not rely on conditional independence tests. (Chickering et al. 1995, Spirtes and Meek 1995) Instead, these are heuristic searches that attempt to maximize a score. We will describe here a heuristic PAG search that attempts to find a PAG with the highest score. One problem with this approach is that because a PAG represents a set of DAG models which may receive different scores (either Bayes Information Criterion, posterior probability, etc.) a PAG cannot be assigned a score by setting its score equal to an arbitrarily chosen DAG that it represents. In the next section we will show how to indirectly assign a score to a PAG.

IV. Mixed Ancestral Graphs (MAGs)

A MAG (or mixed ancestral graph) is a completely oriented PAG for a set of graphs which consists of a single directed acyclic graph $G(\mathbf{O},\mathbf{L})$. (By completely oriented we mean that there are no "o" endpoints on any edge). Some examples of MAGs are shown in Figure 1, where $\mathbf{O} = \{A,B,C,D\}$.

A MAG can also be considered a representation of a set of conditional independence relations among variables in **O** (which in some cases cannot be represented by any DAG containing just variables in **O**; e.g. MAG(G_1(**O**,**L**)) in Figure 1.) A MAG imposes no restrictions on the set of distributions it represents other than the conditional independence relations that it entails. (The class of MAGs is neither a subset nor a superset of other generalizations of DAGs such as chain graphs, cyclic directed graphs, or cyclic chain graphs.)

MAGs have the following useful features:

- DAG G_1 in Figure 1 is an example of a DAG such that as the sample size increases without limit, the difference between the Bayes Information Criterion (BIC) of MAG(G_1,**O**) and the BIC of any DAG G' that contains only variables in **O** increases without limit almost surely. Hence in some cases a maximum likelihood estimate of the MAG parameters is a better estimator of some of the population parameters than the maximum likelihood estimate of any DAG parameters.

- In the large sample limit, for multi-variate normal or discrete distributions, any (possibly latent variable) DAG with a maximum BIC score is represented by the MAG with the highest BIC score among all MAGs.
- There is a three place graphical relation among disjoint sets of vertices (**A** is d-separated from B given **C**) which holds if and only if the MAG entails that **A** is independent of **B** conditional on **C**. D-separation in MAGs is a simple extension of Pearl's d-separation relation (Pearl 1988) defined over DAGs.

If a PAG Ψ represents **Equiv**(G(**O**,**L**)), we say that any MAG that represents graph G(**O**,**L**) is represented by Ψ. For every PAG, there is some MAG that it represents, and every MAG represented by a PAG receives the same BIC score. Thus a PAG can be assigned a score by finding some MAG that it represents, scoring the MAG, and assigning that score to the PAG. It is possible that a PAG represents some non-MAG model that receives a higher BIC score than any MAG represented by the PAG. However, assigning a MAG score to a PAG that represents it has the following desirable property. For any distribution P(**O**), if there is some DAG G that contains **O**, such that for any three disjoint sets of variables **X**, **Y**, **Z** $\subseteq$ **O**, **X** is independent of **Y** given **Z** if and only if **X** is d-separated from **Y** given **Z** in G, then **P**(**O**) is said to be **faithful** to G over **O**. For any multi-variate normal distribution P(**O**), if P(**O**) is faithful to some DAG G over **O**, then in the large sample limit the PAG that represents G receives the highest BIC score among all PAGs.

A. *Parameterizing Gaussian MAGs*

We will describe how a parameterization of a MAG in the multi-variate normal case is an extension of a parameterization of a DAG corresponding to a "structural equation model". (Parameterization and estimation of parameters in the case of discrete variables is an area of current research.)

The variables in a linear structural equation model (SEM) can be divided into two sets, the "error variables" or "error terms," and the substantive variables. Corresponding to each substantive variable X_i is a linear equation with X_i on the left hand side of the equation, and the direct causes of X_i plus the error term ε_i on the right hand side of the equation. Since we have no interest in first moments, without loss of generality each variable can be expressed as a deviation from its mean.

Consider, for example, two SEMs S_1 and S_2 over **X** = $\{X_1, X_2, X_3\}$, where in both SEMs X_1 is a direct cause of X_2. The structural equations in Figure 2 are common to both S_1 and S_2:

$$X_1 = \varepsilon_1$$

$$X_2 = \beta_{21} X_1 + \varepsilon_2$$

$$X_3 = \varepsilon_3$$

Figure 2: Structural Equations for SEMs S_1 and S_2

where β_{21} is a free parameters ranging over real values, and ε_1, ε_2 and ε_3 are error terms. In addition suppose that ε_1, ε_2 and ε_3 are distributed as multivariate normal. In S_1 we will assume that the correlation between each pair of distinct error terms is fixed at zero. The free parameters of S_1 are $\theta = <\beta, \mathbf{P}>$, where β is the set of linear coefficients $\{\beta_{21}\}$ and **P** is the set of variances of the error terms. We will use $\Sigma_{S_1(\theta_1)}$ to denote the covariance matrix parameterized by the vector θ_1 for model S_1. If all the pairs of error terms in a SEM S are uncorrelated, we say S is a SEM with **uncorrelated errors**.

S_2 contains the same structural equations as S_1, but in S_2 we will allow the errors between X_2 and X_3 to be correlated, i.e., we make the correlation between the errors of X_2 and X_3 a free parameter, instead of fixing it at zero, as in S_1. In S_2 the free parameters are $\theta = <\beta$, **P'**>, where β is the set of linear coefficients $\{\beta_{21}\}$ and **P'** is the set of variances of the error terms and the correlation between ε_2 and ε_3. If the correlations between any of the error terms in a SEM are not fixed at zero, we will call it a SEM with **correlated errors**.

It is possible to associate with each SEM with uncorrelated errors a directed graph that represents the causal structure of the model and the form of the linear equations. For example, the directed graph associated with the substantive variables in S_1 is $X_1 \rightarrow X_2$ X_3, because X_1 is the only substantive variable that occurs on the right hand side of the equation for X_2.

It is generally accepted that correlation is to be explained by some form of causal connection. Accordingly if ε_2 and ε_3 are correlated we will assume that either ε_2 causes ε_3, ε_3 causes ε_2, some latent variable causes both ε_2 and ε_3, or some combination of these. We represent the correlated error between ε_2 and ε_3 by introducing a latent variable T that is a common cause of X_2 and X_3. If $\mathbf{O} = \{X_1, X_2, X_3\}$, the MAG for the directed graph associated with S_2 is $X_1 \rightarrow X_2 \leftrightarrow X_3$. The statistical justification for this is provided in Spirtes et al. (1996). It turns out that the set of MAGs is a subset of the set of recursive structural equation models with correlated errors. Hence, there are well known techniques (Bollen, 1992) for estimating and performing statistical tests upon MAG models such as S_2.

B. *The Bayes Information Criterion (BIC) Score of a 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 multi-variate normal structural equation model, for a given sample

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

$$= \text{constant} + \text{Deviance(M)} + \ln(\text{samplesize}) * \text{df}_M$$

where

- θ_{max} is the maximum likelihood estimate of the parameters for model M from sample,
- $\Sigma_{M(\theta_{\max})}$ is the covariance matrix for M when θ takes on its maximum likelihood value θ_{max},
- $\text{L}(\ \Sigma_{M(\theta_{\max})}, \text{sample})$ is the likelihood ratio test statistic of $\Sigma_{M(\theta_{\max})}$,
- df_M is the degrees of freedom of the MAG M.
- Deviance(M) is twice the difference between the unconstrained maximum of the log-likelihood and the maximum taken over M.

(See Raftery, 1993).

V. 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

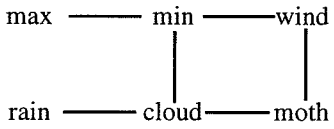
The data as given by Cochran are:

	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:



Dimension	**Deviance**	**Deviance +ln(SS)*dim.**
12	15.66	66.90

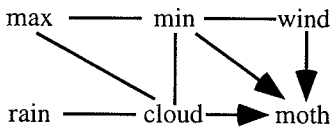
Figure 3: Dempster's Model

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 was greater than 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:



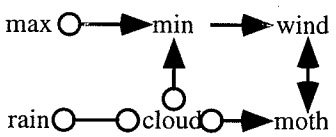
Dimension	Deviance	Deviance +ln(SS)*dim.
14	4.42	64.52

Figure 4: Whittaker's Model[1]

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* ⫫ *wind* | *min*, and does not imply *cloud* ⫫ *wind* | *min*, *moth* whereas the reverse is true of Dempster's model.

C. *FCI Model*

Applying the FCI algorithm to this data resulted in the Partial Ancestral Graph shown in Figure 5. (The structural equation modelling programme EQS, developed by Peter Bentler, was used to fit these models.)



Dimension	Deviance	Deviance +ln(SS)*dim.
12	6.53	57.77

Figure 5: PAG found by FCI search

This PAG 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 PAG. Further, the PAG implies the following structural properties are true of any DAG (possibly with latent variables) which is conditional independence equivalent to the PAG:

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*;

---

[1]When we used the Bentler's EQS programme to fit the model Whittaker reports a devianceof 4.42.

min is an ancestor of *wind*;

there is an unmeasured common cause of *moth* and *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 = 0.39). 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 $max \perp\!\!\!\perp cloud \mid min, wind, rain$. The FCI model manages to accommodate 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.

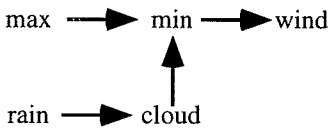


Figure 6: DAG model for the covariates

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 PAG, 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 PAG 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 precede *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*.

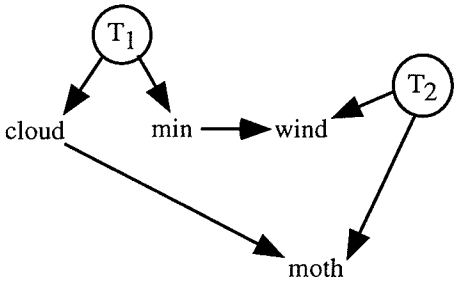


Figure 8: A DAG with latent variables

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.

As described in section IV, it is not possible to score a PAG directly, but only through parameterizing a MAG represented by it. The MAG used for scoring the FCI PAG is shown in Figure 9.

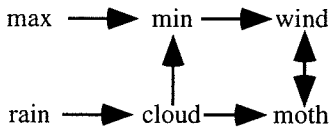


Figure 9: MAG corresponding to PAG in Fig. 5

Note that the same BIC score would be obtained from scoring any other MAG represented by the PAG. 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 PAG with the best BIC score (though it will do so asymptotically). However, it appears that in this example the FCI algorithm did locate the PAG with the best score; a greedy search failed to find a PAG with a higher score. A number of other PAGs, together with the associated deviance and scores are given in Figure 10.

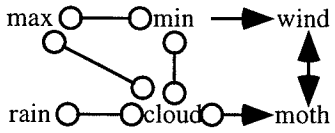
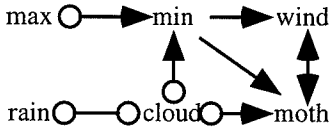
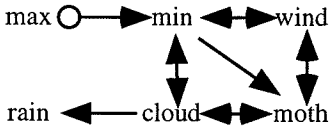
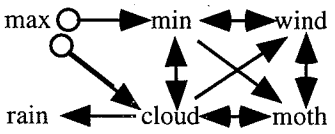
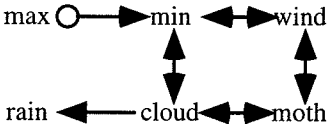
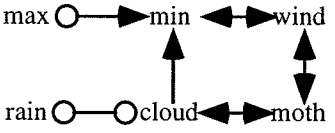
	Dimension	Deviance	Deviance +ln(SS)*dim.
	13	6.50	62.01
	13	4.77	60.28
	13	4.77	59.11
	15	2.26	66.31
	12	11.13	62.37
	12	7.52	58.76

Figure 9: Other PAG Models

VI. *Summary and Future Work*

In this paper have introduced Partial Ancestral Graphs as a representation for equivalence classes of DAGs with latent variables, that captures structural features that are common to all the DAGs in a given equivalence class. We have presented a method for parameterizing the set of Gaussian distributions defined by these conditional independence constraints. This allows a BIC score to be calculated for a PAG. Finally, we have illustrated through an example, that the class of PAG models allow greater flexibility in representing conditional independence, leading to more parsimonious models. Research is currently focused upon

parameterizing MAGs with discrete variables, and developing fast algorithms for transforming PAGs into MAGs.

VII. *Bibliography*

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.

Chickering, D. and Geiger, D. and Heckerman, D. (1995). Learning Bayesian networks: Search methods and experimental results. Preliminary papers of the fifth international workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL, pp. 112-128.

Dempster, A.P. (1972). Covariance Selection. Biometrics, 28, pp. 157-175.

Lauritzen, S., Dawid, A., Larsen, B., Leimer, H., (1990) Independence properties of directed Markov fields, Networks, 20, 491-505.

Pearl, J., (1988). Probabilistic Reasoning in Intelligent Systems, Morgan Kaufman: San Mateo, CA.

Pearl, J. (1995) Causal diagrams for empirical research, Biometrika, 82.

Raftery, A. (1993) Bayesian Model Selection in Structural Equation Models, in Testing Structural Equation Models, ed. by K. Bollen and S. Long, Sage Publications.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. Uncertainty in Artificial Intelligence, Proceedings, 12th Conference, Morgan Kaufman, CA.

Spirtes, P., Glymour, C., and Scheines, R., (1993) Causation, Prediction, and Search, (Springer-Verlag Lecture Notes in Statistics 81, New York).

Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1996). Using D-separation to Calculate Zero Partial Correlations in Linear Models with Correlated Errors, Carnegie Mellon University Technical Report Phil-72.

Spirtes, P., and Meek, C. (1995). Learning Bayesian Networks with Discrete Variables from Data", in Proceedings of The First International Conference on Knowledge Discovery and Data Mining, ed. by Usama M. Fayyad and Ramasamy Uthurusamy, AAI Press, pp. 294-299.

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.