Heuristic Greedy Search Algorithms for Latent Variable Models

Peter Spirtes (Department of Philosophy, Carnegie Mellon Univeristy, ps7z@andrew.cmu.edu),
Thomas Richardson(Department of Statistics, University of Washington) , and Chris Meek (Microsoft)

I. Introduction

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. The DAG in a Bayesian network can be used to represent both causal hypotheses and sets of probability distributions. 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**. (We adopt the convention that sets of variables are capitalized and boldfaced, and individual variables are capitalized and italicized.) Under the statistical interpretation a DAG *G* can be taken to represent a set of all distributions all of which share a set of conditional independence relations that are entailed by satisfying a local directed Markov property (defined below).

Assumptions linking the statistical and causal interpretations of DAG are discussed in Spirtes, Glymour and Scheines (1993). For a particular set of parameters Θ for a DAG *G*, $G(\Theta)$ is a parametric family of distributions. Many familiar parametric models, such as non-recursive structural equation models with uncorrelated errors, factor analytic models, item response models, etc. are special cases of parameterized DAGs. Bayesian networks have proved useful in expert systems, particularly with classification problems (see references in Pearl 1988) and in predicting the effects of interventions into given causal systems (Spirtes et al. 1993 and Pearl 1995).

There has recently been significant progress in the development of algorithms for learning the 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, to classify, and to predict the effects of interventions in causal systems.

II. Directed Acyclic Graphs (DAGs)

A **directed acyclic graph** (DAG) *G* with a set of vertices **V** can be given two distinct interpretations. (We place sets of variables and defined terms in boldface.) On the one hand, such graphs can be used to represent causal relations between variables, where an edge from A to B in *G* means that A is a direct cause of B relative to **V**. A **causal graph** is a DAG given such an interpretation.

On the other hand, a DAG with a set of vertices **V** can also represent a set of probability measures over **V**. 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*.

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). Variants of probabilistic DAG models were introduced in the 1980's in Pearl (1988) among others. (See Pearl 1988 for references.)

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 $\mathbf{O}$ is the set of observed variables and $\mathbf{L}$ 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 $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ included in $\mathbf{O}$, $G_1(\mathbf{O},\mathbf{L})$ entails $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on $\mathbf{Z}$ if and only if $G_2(\mathbf{O},\mathbf{L})$ entails $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on $\mathbf{Z}$. 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 subsets of $\mathbf{Equiv}(G(\mathbf{O},\mathbf{L}))$. A PAG is an extended graph consisting of a set of vertices $\mathbf{V}$, 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 an 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. 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 more different kinds 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 $*\!\rightarrow$ B represents either A $\rightarrow$ B, A $\leftrightarrow$ B, or A o$\rightarrow$ B.

Partial Ancestral Graphs (PAGs)

If **G** is a set of directed acyclic graphs included in $\mathbf{Equiv}(G(\mathbf{O}, \mathbf{L}))$, Ψ 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 $\mathbf{O}\backslash\{A,B\}$.

(ii) If there is an edge in Ψ out of A, i.e. A $\rightarrow$ 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$*\!\rightarrow$ B, then in every DAG in **G**, B is **not** an ancestor of A.

(iv) If there is an underlining A*—$\underline{\text{*B*}}$—*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 $\mathbf{O}$ = {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 $\mathbf{Equiv}(G(\mathbf{O}, \mathbf{L}))$ guarantees that if one directed acyclic graph in $\mathbf{Equiv}(G(\mathbf{O}, \mathbf{L}))$ does not entail that A and B are independent conditional on any subset of $\mathbf{O}\backslash\{A,B\}$, then all directed acyclic graphs in $\mathbf{Equiv}(G(\mathbf{O}, \mathbf{L}))$ do not entail that A and B are independent conditional on any subset of $\mathbf{O}\backslash\{A,B\}$.

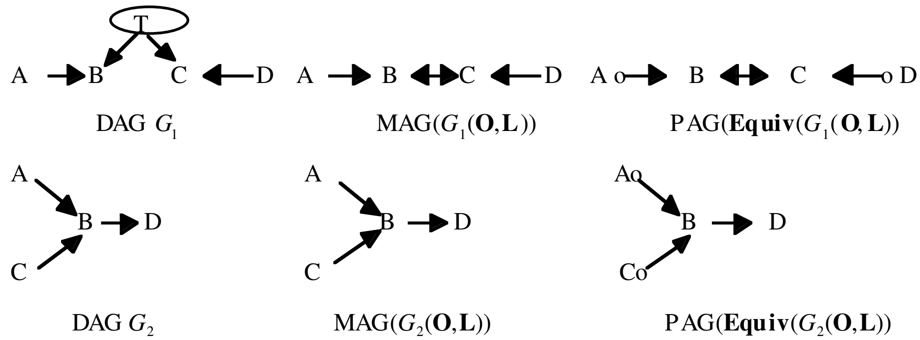

Figure 1

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** $\cup$ **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** = $\emptyset$.

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 consistently estimate the effect of the intervention, even without knowing which causal models 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 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 directly. 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 $\mathbf{O} = \{A,B,C,D\}$.

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 the graph $G(\mathbf{O},\mathbf{L})$ that a MAG represents is included in the PAG that represents **Equiv**($G(\mathbf{O},\mathbf{L})$), then we say that the PAG represents the MAG. 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 $P(\mathbf{O})$, if there is some DAG G that contains **O**, such that for any three disjoint sets of variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{O}$, **X** is independent of **Y** given **Z** if and only if **X** is d-separated from **Y** given **Z** in G, then $\mathbf{P}(\mathbf{O})$ is said to be **faithful** to G over **O**. For any multi-variate normal distribution $P(\mathbf{O})$, if $P(\mathbf{O})$ 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 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($G_1(\mathbf{O},\mathbf{L})$) 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.)

A. Parameterizing 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 somewhat more difficult.)

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 $\mathbf{X} = \{X_1, X_2, X_3\}$, 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 $\varepsilon_1, \varepsilon_2$ and ε_3 are error terms. In addition suppose that $\varepsilon_1, \varepsilon_2$ 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 $\theta = \langle\beta, \mathbf{P}\rangle$, where β is the set of linear coefficients $\{\beta_{21}\}$ and $\mathbf{P}$ 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, and occasionally leave out the model subscript if the context makes it clear which model is being referred to. 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 X_1. In X_2 the free parameters are $\theta = \langle\beta, \mathbf{P'}\rangle$, where β is the set of linear coefficients $\{\beta_{21}\}$ and $\mathbf{P'}$ 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 $X_1 \rightarrow X_2 \quad X_3$, 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 standard statistical packages such as LISREL or EQS which can estimate and perform 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 the posterior distribution. In the case of a multi-variate normal structural equation model, for a given sample,

$$\mathrm{BIC}(M, \text{sample}) = L(\Sigma_{M(\theta_{max})}) - \ln(\text{samplesize} * \text{number of variables}) * df_M, \text{ where}$$

- θ_{max} is the maximum likelihood estimate of the parameters for model M,
- $\Sigma_{M(\theta_{max})}$ is the covariance matrix for M when Θ takes on its maximum likelihood value θ_{max},
- $L(\Sigma_{M(\theta_{max})})$ is the likelihood ratio test statistic of $\Sigma_{M(\theta_{max})}$,
- df_M is the degrees of freedom of the MAG M.

(See Raftery, 1993). Each of these quantities can be calculated by standard statistical packages such as LISREL or EQS.

C. Greedy BIC MAG Search

A greedy search among MAGs is given as input a MAG to start with (possibly a MAG with no edges). At each stage, the algorithm takes the MAG M it has constructed thus far and calculates the score of each MAG resulting from a one edge addition (directed or bi-directed) to M, removal of one edge (directed or bi-directed) from M, or reversal of one directed edge in M. If none of these changes improves the BIC score of M, the algorithm halts and outputs M. Otherwise the change that most improves the BIC score is made to M and the process is repeated.

Even at large sample sizes, this search suffers from the following problem. At each stage, there may be many MAGs that receive the same BIC score, and the algorithm arbitrarily chooses one of them. While two MAGs M and M' may receive the same BIC score, the one edge modification to M may receive a much higher BIC score than the same one edge modification to M'. Thus if the search halts when it cannot improve the score, it may halt at M', which is a local rather that a global maximum.

V. Greedy BIC PAG Search

A greedy PAG search solves some of the problems associated with a greedy MAG search. First, there are many fewer MAGs than PAGs. Second, MAGs that are score-equivalent in the sense that they receive the same BIC score for every data set will all be represented by the same PAG, and no two PAGs are score-equivalent. Hence the search does not suffer from the problem of having to choose arbitrarily from many score-equivalent alternatives. The search is described below and illustrated in Figure 3. It is basically a latent variable versions of a search devised by C. Meek and described in Spirtes and Meek (1995).

```
procedure GBPS(PAG; data);
begin
  MAG:=PAG-to-MAG(PAG);
  current-score:=score(MAG,data);
  max-score:=current-score;
  while max-score <= current-score do
   begin
    new-PAG:=add-best-edge-to-PAG(PAG);
    MAG:=PAG-to-MAG(new-PAG);
    current-score:=score(MAG,data);
    if current-score > max-score then
     begin
      max-score:=current-score;
       PAG:=new-PAG;
     end;
  end;
  current-score:=max-score;
  while max-score <= current-score do
   begin
    new-PAG:=remove-worst-edge-in-PAG(PAG);
    MAG:=PAG-to-MAG(new-PAG);
    current-score:=score(MAG,data);
    if current-score > max-score then
     begin
      max-score:=current-score;
       PAG:=new-PAG;
     end;
  end;
  return(PAG);
end;
```

The search starts with some initial PAG. This could come from background knowledge, another search procedure such as FCI, or could simply be a PAG with no edges. The PAG is then turned into a MAG in order to assign a score to it. The search then looks for the single best edge to add to the initial PAG. In the example of Figure 3 there are four single edge PAG extensions of the initial PAG. Each of these four extensions is turned into a MAG in order to score it. The MAG with the best score is chosen, and turned back into its corresponding PAG. These steps are then repeated until the score cannot be improved. At this stage the search then removes edges until the score can no longer be improved.

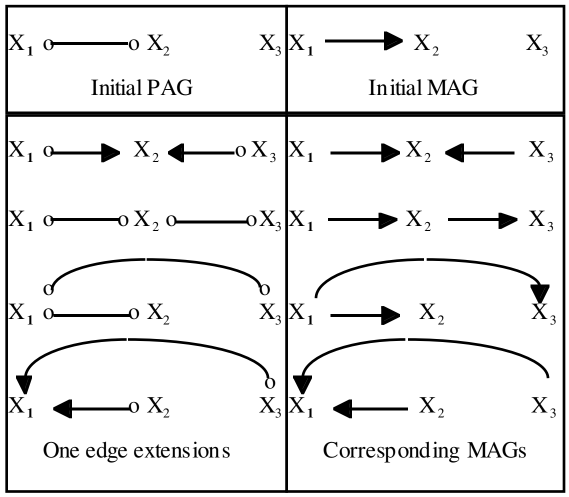

Figure 3

We conjecture that this search is asymptotically correct, as long as the distribution is the marginal of a distribution faithful to some directed acyclic graph. It is worst case exponential in the number of vertices because of the remove-worst-edge-in-PAG step. In addition, we do not know the complexity of the PAG-to-MAG step, because we do not know how much back-tracking may be needed in order to turn a PAG into a MAG. We do not currently know the number of variables that this kind of search can feasibly be performed on. The current implementation is not practical for 30 variables, but could be greatly speeded up.

VI. Simulation Study

As a very preliminary simulation study, we chose two graphs G_1 and G_2 with latent variables (Figure 4), and for sample sizes 2500, 1000, 500, and 250 generated 5 pseudo-rendom samples from them. The error variables were standard normal, and the linear coefficients were between 0.5 and 1.5, and did not vary with sample size or sample. The input to the algorithm is the data, and the output is a PAG. Because determining whether X is an ancestor of Y is important for predicting the effects of interventions on X, we measure the performance of the algorithm by counting for how many ordered pairs of variables <X,Y> the output PAG implies that X is an ancestor of Y (#a in Table 1, averaged over the 5 samples at a given sample size), and what percentage of the time the ancestor implication is correct in the graph that generated the data (%ac, averaged over the 5 samples at a given sample size). We construct similar measured for non-ancestor relations (#na and %nac respectively). In Table 1, size represents the sample size. In G_1, in 20% of the ordered pairs of distinct measured variables <X,Y>, X is an ancestor of Y; in G_2, in 30% of the ordered pairs of distinct measured variables <X,Y>, X is an ancestor of Y. In the case of large sample sizes and sparse graphs, with perfectly normal data, and only a few latent variables, the algorithm performs quite well (see Table 1). However, we expect the algorithm's performance to be a function of the edge coefficients, how many vertices each vertex in the graph is adjacent to, the sample size, the number and strength of latent variables, the amount of selection bias, and deviations from normality. In order to evaluate the algorithm's performance much more extensive simulation tests are needed, as well as applications to real data.

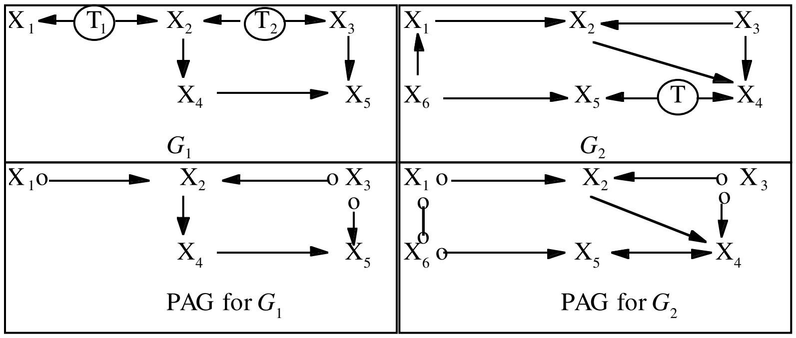

Figure 4

	*G*1					*G*2				
size	2500	1000	500	250	100	2500	1000	500	250	100
#a	3	3	3	3	3	1	1	1	.6	.6
%ac	100	100	100	100	100	100	100	20	75	100
#na	11	11	11	11	11	19	19	19	12.4	7.4
%nac	100	100	100	100	100	100	100	100	98.3	94.6

Table 1

Bibliography

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.

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.

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