Discovering Causal Relations Among Latent Variables in Directed Acyclic Graphical Models

by

Peter Spirtes

February 1996

Report CMU-PHIL-69

Carnegie Mellon

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Discovering Causal Relations Among Latent Variables in Directed Acyclic Graphical Models[1]

by

Peter Spirtes
Department of Philosophy
Carnegie-Mellon University
e-mail: ps7z@andrew.cmu

1. Introduction

Many theories suppose there are variables that have not been measured but that influence measured variables. In studies in econometrics, psychometrics, sociology and elsewhere the principal aim may be to uncover the causal relations among such "latent" variables. In such cases the measuring instruments are often designed with fairly definite ideas as to which measured items are caused by which unmeasured variables. Survey questionnaires may involve hundreds of items, and the very number of variables is ordinarily an impediment to drawing useful conclusions about structure.

[1] I thank Richard Scheines and Clark Glymour for many helpful discussions. Research for this paper was supported by the National Science Foundation through grant 9102169 and the Navy Personnel Research and Development Center and the Office of Naval Research through contract number N00014-93-1-0568.

In this paper I propose an algorithm for constructing a set of causal structures from sample data, given some background knowledge about which measured variables are indicators of which latent variables. The algorithm is an extension of work described in Chapter 10 of Spirtes, Glymour and Scheines(1993), and Scheines(1993). Under a set of assumptions stated in the following sections, the set of alternative models output from population statistics is certain to contain the true model. The algorithm performs statistical tests on samples of measured variables in order to make decisions about conditional independence relations[2] among the latent variables in the population. These decisions about conditional independence relations in the population are then used to construct (using the PC or FCI algorithms described in Spirtes, Glymour, and Scheines, 1993) a set of causal structures compatible with the conditional independence relations judged to hold in the population.

Section 2 describes the problem more precisely. Section 3 lists the assumptions made and outlines the algorithm. Sections 4, 5, and 6 describe the three portions of the algorithm in more detail. Finally, section 7 describes a simulation study of an application of the algorithm to linear models. The proofs of the theorems are in the Appendix (except for Theorem 1, which is a slight modification of a theorem proved in Spirtes, Glymour and Scheines(1993) and Scheines(1993).)

2. Directed Acyclic Graph (DAG) Models

Factor analysis models, path models with jointly independent errors, recursive linear structural equation models with jointly independent errors, and various kinds of latent variable models including latent class analysis models, latent profile analysis models, and latent trait analysis models, are all instances of DAG models.

A **directed graph**[3] is an ordered pair of a finite set of vertices **V**, and a set of directed edges **E**. A directed edge from *A* to*B* is an ordered pair of distinct vertices $<A,B>$ in **V** in which *A* is the **tail** of the edge and *B* is the **head**; the edge is **out of** *A* and **into***B*, and *A* is **parent** of *B* and *B* is a **child** of *A*.*A* and *B* are **adjacent** if and only if there is a directed edge from *A* to *B* or from *B* to *A*. A sequence of vertices $<V_1,...,V_n>$ in *G* is an **undirected path** between V_1 and V_n if and only if for $1 \leq i < n$, V_i and V_{i+1} are adjacent in *G*. A path *U* is **acyclic** if no vertex in the path occurs more than once. A sequence of

[2] "Conditional independence relations" refers to both independence relations and conditional independence relations.

[3] Sets of variables and terms being defined are placed in boldface, and individual variables in italics.

vertices $<V_1,...,V_n>$ in G is a **directed path** from V_1 to V_n if and only if for $1 \leq i < n$, there is a directed edge from V_i to V_{i+1} in G. If there is an acyclic directed path from A to B or $B = A$ then A is an **ancestor** of B, and B is a **descendant** of A. A directed graph is **acyclic** if and only if it contains no directed cyclic paths.

A **directed acyclic graph** (DAG) G with a set of vertices **V** can be given both a causal interpretation and a statistical interpretation. A set **V** of variables is **causally sufficient** in a given population if and only if every common cause of a variable in **V** is also in **V**. A DAG can be used to represent causal relationships between causally sufficient sets of variables; under this interpretation a DAG will be called a **causal DAG**. If **V** is a causally sufficient set of variables, there is an edge from A to B in a causal DAG G with variables **V** if and only if A is a direct cause of B relative to **V**. 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 DAG G with vertices **V** if and only if for every W in **V**, W is independent of $\mathbf{V} \backslash (\mathbf{Descendants}(W,G) \cup \mathbf{Parents}(W,G))$ given $\mathbf{Parents}(W,G)$, where $\mathbf{Parents}(W,G)$ is the set of parents of W in G, and $\mathbf{Descendants}(W,G)$ is the set of descendants of W in G. A DAG G **represents** the set of probability measures which satisfy the local directed Markov property for G. If a conditional independence relation is true in every probability measure that satisfies the local directed Markov property for DAG G, say that G **entails** the conditional independence relation. If every conditional independence relation true in a probability measure P is entailed by DAG G, say that P is **faithful** to G. The use of DAGs to simultaneously represent a set of causal hypotheses and a family of probability measures extends back to the path diagrams introduced by Sewell Wright(1934). Variants of DAG models were introduced in the 1980's in Wermuth(1980), Wermuth and Lauritzen(1983), Kiiveri, Speed, and Carlin(1984), Kiiveri and Speed(1982), and Pearl(1988). For simplicity, it will always be assumed that the probability measures represented have densities. Then if a probability measure with density $f(\mathbf{V})$ satisfies the local directed Markov property for DAG G, there is a factorization of $f(\mathbf{V})$ of the form:

$$f(\mathbf{V}) = \prod_{V \in \mathbf{V}} f(V | \mathbf{Parents}(\mathbf{V},G))$$

Recursive linear structural equation model, or RSEMs (adapting the terminology in Bollen, 1989) with jointly independent error terms can be represented as DAG models in which each variable is a linear function of its parent in the DAG and an error term with a non-zero variance. The error terms are not represented in the DAG. Call the assignment of the coefficients and the variances of the exogenous variables a linear parameterization of a DAG. The linear coefficients and the variances of the exogenous variables determine the correlation matrix. In order to fully specify the probability measure, it is necessary to specify the joint probability measure over the exogenous variables. It will be assumed that the error terms are jointly independent for a causally sufficient set of variables. Again, this assumption does not entail that any given set of measured variables are jointly independent; this will not be the case if the measured variables are not causally sufficient. Rather, the assumption is that the probability measure over the measured variables is the marginal of a probability measure with jointly independent errors.

Latent class analysis model, described in Bartholomew(1987) are special cases of DAG models in which all of the variables are discrete. In a discrete DAG model all of the variables are discrete variables with a finite number of categories, and the joint probability measure can be factored into the form:

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V | \mathbf{Parents}(\mathbf{V}, G))$$

A discrete parameterization of a DAG G assigns values to $P(V|\mathbf{Parents}(V))$ for each V in the DAG.

The goal is to infer causal relationships among a set of latent variables **L**, given a set of measured variables **M** that are indicators of variables in **L**. It is assumed that there is sample data for the measured variables, and from background knowledge the user specifies which variables in **M** are indicators of which variables in **L**. For example the variables in **L** might be various psychological attributes, and the variables in **M** answers to test questions intended to measure those attributes. $\mathbf{M} \cup \mathbf{L}$ may not be a causally sufficient set of variables. However, it will always be assumed that there is some set of variables **O** such that $\mathbf{M} \cup \mathbf{L} \cup \mathbf{O}$ is causally sufficient. It is also assumed that if the user specifies that a variable M in **M** is an indicator of a variable L in **L**, then M is a child of L in the causal graph (i.e. there is a directed edge from L to M.) However, it is not assumed that M is directly causally connected *only* to the variable in **L** that the user has specified it is an indicator of, i.e. M may also be the child of some other variable in $\mathbf{M} \cup \mathbf{L} \cup \mathbf{O}$.

The DAG in figure 1 will be used to illustrate the algorithm.

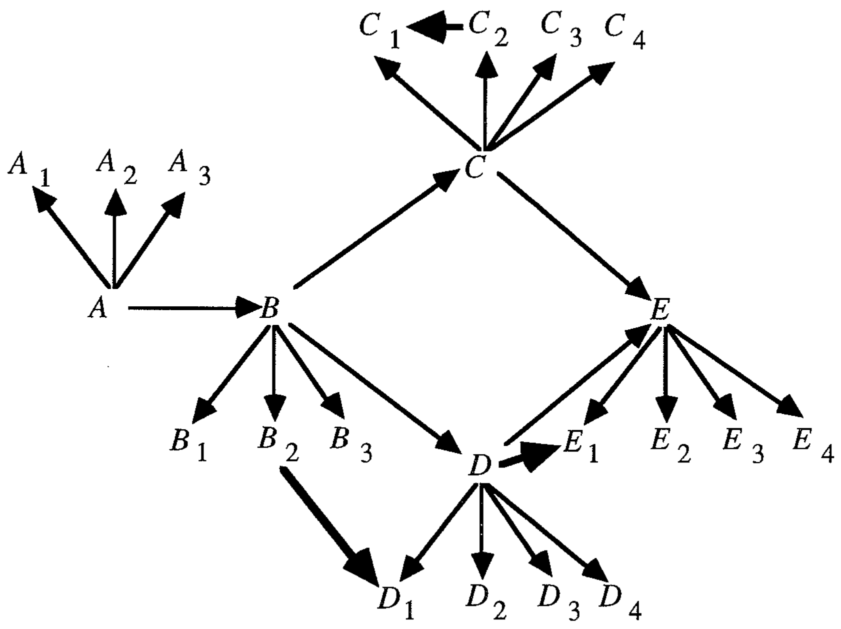


Figure 1: True Causal DAG *G*

Suppose that **M** = $\{A_1,A_2,A_3,B_1,B_2,B_3,C_1,C_2,C_3,C_4,D_1,D_2,D_3,D_4,E_1,E_2,E_3,E_4\}$, **L** = $\{A,B,C,D,E\}$, and **M** ∪ **L** is causally sufficient. Suppose that the user specifies that each variable in **M** is indicator of the variable with the corresponding name in **L** (e.g. A_1, A_2, and A_3 are indicators of A). Each variable in **M** is a child of some variable in **L**. In general, **Measured**(L) is the subset of variables in **M** that the user has specified as indicators of L, and *Latent*(M) is the latent variable specified by the user as the latent that M is an indicator of. For example, in figure 1, **Measured**(A) = $\{A_1,A_2,A_3\}$ and *Latent*(A_1) = A. Call the subgraph containing only the variables in **L** ∪ **O** and the edges between those variables the **structural graph.** The graph with the edges between variables in **L** ∪ **O** removed is called the **measurement graph.**

In contrast to the previous case, suppose instead that the set of measured variables **M** = $\{A_1,A_2,A_3,C_1,C_2,C_3,C_4,D_1,D_2,D_3,D_4,E_1,E_2,E_3,E_4\}$, i.e. the variables $\{B_1,B_2,B_3\}$ were not measured. In this case **L** = $\{A,C,D,E\}$ (because no variable in **M** is an indicator of B) and **M** ∪ **L** is not causally sufficient. However, if **O** = $\{B\}$, then **M** ∪ **L** ∪ **O** is causally sufficient.

It is not assumed that a variable M in **M** is directly causally connected only to *Latent*(M). For example, C_1 is a child of C_2, both of which are indicators of C. D_1 is a child of B_2, which is an indicator of another latent variable, B. Finally, E_1 is a child of another latent variable, D. The edges between measured variables and variables that they are not intended to be indicators of are in boldface in figure 1.

3. Outline of the Algorithm

In this section, the inputs, outputs, and basic structure of the algorithm are described.

3.1 The Input

The input to the algorithm is:

1. sample data for a set of measured variables **M**;
2. an initial measurement DAG which is a subgraph of the true measurement DAG, and in which each measured variable is listed as an indicator of exactly one latent variable in **L**, and each latent variable in **L** has at least one measured indicator;
3. an assumption about the family of distributions that the true model lies in, either discrete (including a specification of the number of categories for each latent variable), or linear;
4. optional background knowledge (such as time order) constraining the causal relationships among the latents;
5. an optional assumption about the causal sufficiency of the latent variables;
6. a significance level for the statistical tests to be performed.

For the example of figure 1, an example of an initial measurement DAG is shown in figure 2.

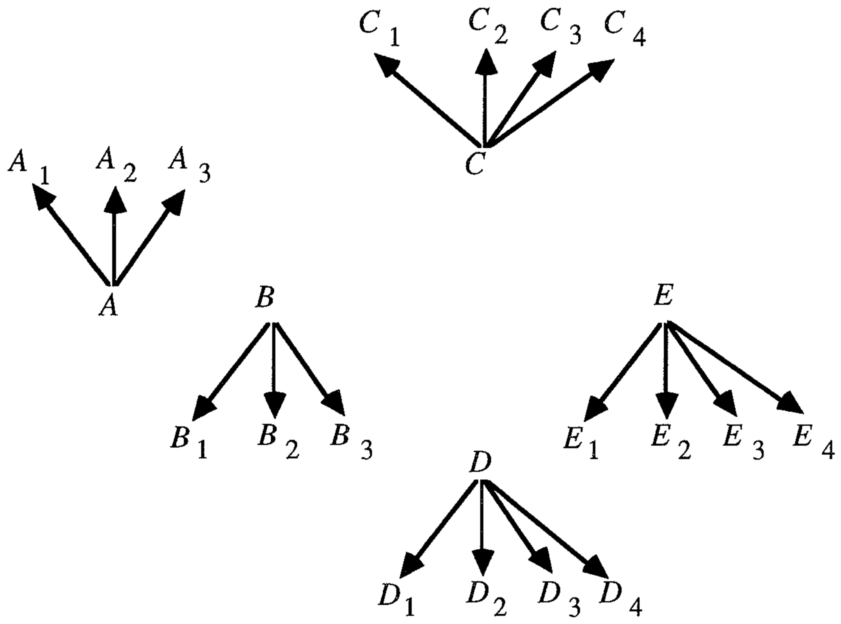


Figure 2: Initial Measurement DAG

3.2 The Output

The algorithm can be run either under the assumption that the latent variables are causally sufficient, or that the latent variables are not causally sufficient. In either case, the output is a set of DAGs. However, the conventions by which sets of DAGs are represented is much more complicated in the latter case. Hence, this section will describe the representation conventions only for the former case. The representation conventions for the latter case are described in Chapter 6 of Spirtes, Glymour, and Scheines(1993).

In order to describe the output of the algorithm the following definitions are needed. Two DAGs with the same set of variables are **faithfully indistinguishable** if and only if they entail the same set of conditional independence relations. The **faithful indistinguishability class** of a DAG *G* is the set of DAGs faithfully indistinguishable from *L*. In some, but not all cases, the **faithful indistinguishability class** of a DAG *G* contains a single DAG.

If causal sufficiency of the latent variables is assumed, the algorithm outputs a **pattern** containing the variables in **L** (see Verma and Pearl, 1990) that represents the subset of a faithful indistinguishability class of DAGs compatible with the optional background knowledge about the causal structure. A pattern is a graphical object that may contain both directed edges and undirected edges. In a DAG or a pattern, *B* is an **unshielded**

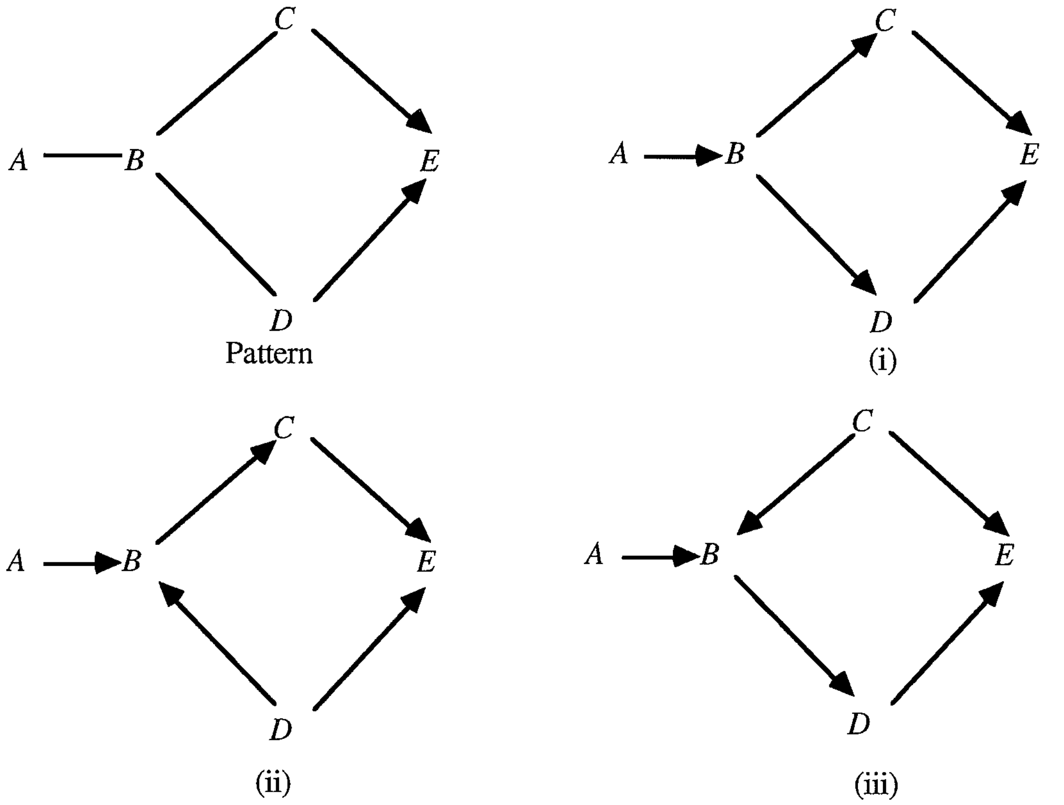
collider on a path U containing $<A,B,C>$ if and only if U contains edges A -> B and C -> B, and A and C are not adjacent in G. A DAG G is in the set of graphs represented by Π if and only if:

(i) G has the same adjacency relations as Π;

(ii) if the edge between A and B is oriented A -> B in Π, then it is oriented A -> B in G;

(iii) if Y is an unshielded collider on the path $<X,Y,Z>$ in G then Y is an unshielded collider on $<X,Y,Z>$ in Π.

Figure 3 illustrates a pattern and the three DAGs that it represents.



Pattern

(i)

(ii)

(iii)

Figure 3: Pattern Representing Faithful Indistinguishability Class of DAG G

Given that figure 1 shows the true causal DAG, and figure 2 the initial measurement DAG, the pattern in figure 3 is the output of the algorithm.

3.3 Outline of the Algorithm

The basic outline of the algorithm is as follows:

1. Given a set of measured variables, find a subset of the measured variables which are almost pure (which, roughly speaking, means that in the true measurement DAG for that subset of measured variables, each measured variable in the subset is not directly causally related to any variables except the latent variable the user specifies that it is an indicator of.) A more precise account of this step and the exact definition of almost purity are given in section 5.

2. Give the sample data for the subset of measured variables chosen in step 1 to the PC algorithm (described in section 4) if the assumption of causal sufficiency of the latent variables is made, or to the FCI algorithm, if it is not.

   a. The PC or FCI algorithms test various conditional independence relations among the latent variables using the tests described in section 6. (In the case of linear parameterizations of DAGs, the algorithm uses tests of zero partial correlations instead of tests of conditional independence, even if the probability measure is not normal.)

The point of step 1 is that is that the subset of variables that it selects have a known measurement DAG. Then using the known measurement DAG, it is possible to detect features of the causal relations between the latent variables.

3.4. The Correctness of the Algorithm

In the linear case, in the large sample limit, the set of DAGs represented by the output of the algorithm contains the true structural DAG *G* with probability 1 under the following assumptions:

1. the user input is correct;
2. the Causal Markov Condition holds: if *G* is a causal DAG with causally sufficient vertex set **V** and *P* is a probability measure over the vertices in **V** generated by the causal structure represented by *G* then *G* and *P* satisfy the Causal Markov Condition if and only if for every *W* in **V**, *W* is independent of **V\\(Descendants**(*W*) ∪ **Parents**(*W*)) given **Parents**(*W*).
3. the Causal Faithfulness condition holds: if *G* is a causal DAG over a causally sufficient set of variables **V**, and *P* a probability measure generated by *G* then <*G*, *P*> satisfies the Faithfulness Condition if and only if every conditional independence relation true in *P* is entailed by the Causal Markov Condition applied to *G*;

4. the true caual DAG has a non-zero prior probability, and for each DAG, the prior distribution over the linear parameters for that DAG (the linear coefficients and the variances of the error variables) is absolutely continuous with Lebesgue measure;
5. after step 1 of the algorithm, each latent variable has at least three measured indicators.

The proof of this follows from the theorems stating the correctness of the individual steps in the algorithm given in the following sections.

The correctness of the tests used in step 2a require that the output of step 1 be a subset of measured variables with an almout pure measurement DAG. In the discrete case, while there are necessary conditions for a subset of variables to be almost pure (stated in section 5), I do not know of any sufficient conditions. So the assumptions needed to prove the corresponding correctness result for the discrete case are the same as above, except that assumptions 4 and 5 are replaced with the assumption that the subset of variables output by step 1 is almost pure.

The meaning and justification of assumptions 2 and 3 are discussed more fully in Spirtes, Glymour and Scheines(1993). However, if one makes these assumptions then the probability measure of data generated by a causal process represented by causal DAG *G* is a member of the set of probability measures represented by DAG *G*.

In assuming the Causal Markov and Faithfulness Conditions it is not being assumed that any given set of *measured* variables is causally sufficient. It is assumed that the probability measure over a set of measured variables is the *marginal* of a probability measure faithful to the causal graph that generated it. In figure 1, $\{A_1,A_2,A_3\}$ is not a causally sufficient set of variables, but it can be expanded to the causally sufficient set of variables $\{A_1,A_2,A_3,A\}$, and it is assumed that the probability measure over $\{A_1,A_2,A_3,A\}$ is faithful to the subgraph of *G* containing just those variables.

Under the assumption that the set **L** of latent variables is causally sufficient, for models with causal graphs represented by DAGs of a fixed maximum order (the order of a vertex is the number of other vertices it is adjacent to) the number of tests of conditional independence relations the algorithm must perform is polynomial in the number of latent vertices.

4. The PC and FCI Algorithms

The PC algorithm takes as input optional background information about the causal structure, and uses tests of conditional independence relations among a set of variables **V** to construct a pattern containing variables **V** that represents the subset of a faithful indistinguishability class compatible with the user entered background knowledge. The current implementation (see Scheines, Spirtes, Meek, and Glymour forthcoming) uses the sample data for the measured variables to test conditional independence relations among the measured variables. This version of the algorithm could not be applied to the problem described here, because the measured variables are not causally sufficient. However, the next two sections describe how to use sample data for the measured variables to test conditional independence among the latent variables. If the set of latent variables is causally sufficient, the PC algorithm can then use these tests to construct a pattern containing the latent variables.

Like the PC algorithm, the FCI takes as input optional background information about the causal structure, and uses tests of conditional independence relations to construct a set of causal DAGs. The current implementation (see Scheines, Spirtes, Meek, and Glymour forthcoming) also uses the sample data for the measured variables to test conditional independence relations among the measured variables. Unlike the PC algorithm, the FCI algorithm does not assume that the measured variables are causally sufficient, so it could be applied directly to the sample data for the measured variables. However, while the output of the FCI algorithm is in some cases informative about the existence and relationships among latent variables, in the kinds of models considered here, the output would be correct but extremely uninformative in the sense that it places very few constraints on the causal relationships among the latent variables. The methods for testing conditional independence among the latent variables using the sample data among the measured variables described in the next two sections make the output of the FCI algorithm much more informative for the kinds of models considered here.

Given the first three assumptions listed in section 3.4, and correct statistical decisions about conditional independence relations in the population, the output of both the FCI and PC algorithms are correct in the sense that the true causal DAG is a member of the set of DAGs represented by their output.

5. Purification

Before testing for conditional independence relations among the latent variables, a search is performed for a submodel of the measurement DAG of $\mathbf{M} \cup \mathbf{L}$ in which the measured variables bear a particularly simple relationship to the latent variables. The basic idea is to find a subset $\mathbf{M'}$ of $\mathbf{M}$ such that every variable M in $\mathbf{M'}$ is a child of no other variable in $\mathbf{M'} \cup \mathbf{L} \cup \mathbf{O}$ except for *Latent*(M). In that case we say that the measurement DAG is **pure**. Finding such a subset is called **purification** of the measurement DAG. (Anderson and Gerbing, 1982, Anderson and Gerbing, 1988, Spirtes, Glymour, and Scheines, 1993, and Scheines, 1993 all discuss purification in the linear case.) By purifying a measurement DAG it is possible to determine the form of the measurement DAG without knowing the structural model. Then, using the known form of the measurement DAG, it is possible to determine the conditional independence relations among the latent variables.

5.1 The Input

The input to the algorithm is:

1. sample data for a set of measured variables **M**;
2. an initial measurement DAG that is a subgraph of the true measurement DAG, and in which each measured variable is listed as an indicator of exactly one latent variable in **L**, and each latent variable in **L** has at least one measured indicator;
3. an assumption about the family of distributions that the true model lies in, either discrete (including a specification of the number of categories for each latent variable), or linear;
4. a significance level for the statistical tests it performs.

5.2 The Output

The output of the algorithm is a pure measurement DAG that contains a subset of the variables in **M**. For example, if figure 1 is the true causal DAG, (where $\mathbf{L} = \{A,B,C,D,E\}$, $\mathbf{M} = \{A_1,A_2,A_3,B_1,B_2,B_3,C_1,C_2,C_3,C_4,D_1,D_2,D_3,D_4,E_1,E_2,E_3,E_4\}$ and $\mathbf{O} = \varnothing$), the true measurement DAG is not pure, because D_1 is is a child of B_2 as well as D; C_1 is a child of C_2 as well as C; and E_1 is a child of D as well as E. However, the true measurement DAG for the subset of **M** that does not contain D_1, C_2, or E_1 is pure. If the initial measurement DAG is the DAG in figure 2, the output of the algorithm is the pure measurement DAG shown in figure 4.

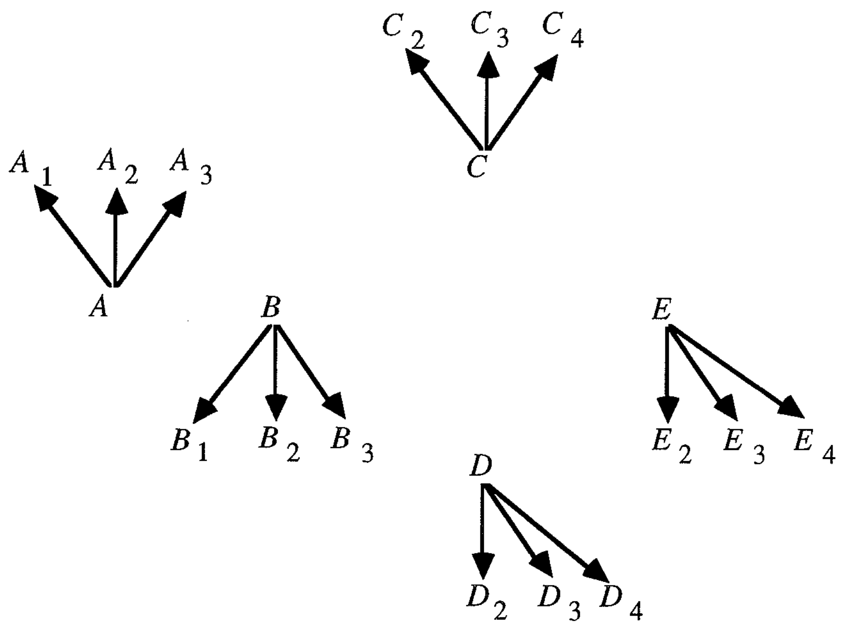


Figure 4: Purified Measurement DAG

There are other subsets of **M** that also have pure measurement DAGs. For example, the subset of **M** that does not contain D_2, C_2, or E_1 has a pure measurement DAG.

5.3 Outline of the Algorithm

A **one factor** DAG of a set of variables **X** is a DAG with a set of vertices $\mathbf{X} \cup \{L\}$, and edges from L to each member of **X**. The following procedure is a generalization of a procedure described in Scheines(1993). An $n \times 1$ set of variables consists of n variables that are indicators of one latent and one variable that is an indicator of another latent. The following algorithm performs a number of statistical tests of one factor models. If the true causal DAG is assumed to be a linear RSEM, then the statistical tests are done under the assumption that each one factor model is a linear RSEM. If the true causal DAG is assumed to be discrete (where for each latent variable, the number of categories is specified) the statistical test for a one factor model of an $n \times 1$ set of variables is done under the assumption that the one factor model is discrete. In addition, the number of categories for the latent variable in the one factor model is set equal to the number of categories for the latent variable with n indicators in the set of measured variables.

For each variable M in **M**, let $score(M)$ be the number of $n \times 1$ sets of variables containing M such that the one factor model of M fails a statistical test, and no one factor model of a subset of the $n \times 1$ set of variables is overidentified. The simplest version of the algorithm is presented below.

```
set $\mathbf{M'} = \mathbf{M}$;
repeat
    for each variable $M$ in $\mathbf{M'}$, calculate $score(M)$;
    for some variable $M$ in $\mathbf{M'}$ such that no other variable in $\mathbf{M'}$ has a higher score,
    remove $M$ from $\mathbf{M'}$;
until for each $M$ in $\mathbf{M'}$, $score(M) = 0$.
```

The scoring function can also be modified to take into account how badly a one-factor model fails a test, as well as how many tests are failed. The algorithm could also be halted if at some point the number of one factor models failing tests is small, and removing many more variables fails to change that number much.

Suppose that the data was generated by a linear parameterization of the DAG in figure 1. The user inputs a pure measurement DAG shown in figure 2. The output of the purification algorithm is shown in figure 4. No matter what the causal relationships between the latents, a DAG with the measurement DAG shown in figure 2 is not entailed to fit the data because in the true DAG there are edges from C_2 to C_1, from B_2 to D_1, and from D to E_1. For example, as explained below, any 3×1 foursome that contains an indicator of D and any three indicators of E including E_1, is not entailed to have a one factor linear model. On the other hand any 3×1 foursome that contains an indicator of D and any three indicators of E excluding E_1, is entailed to have a one factor linear model. So E_1 occurs in more foursomes of variables that fail one factor tests than E_2, E_3, or E_4, and at some point in the algorithm, E_1 is be a better candidate for removal than any other variable that has not yet been removed from $\mathbf{M'}$.

It is possible to test for the existence of one factor linear models in a variety of different ways. See Bartholomew(1985) for a summary of methods relevant to such tests. Relevant tests of models are also discussed in Akaike(1983), Schwartz(1978), and Bozdogan and Ramirez(1986) for normally distributed variables, and Amemiya and Anderson(1985), Fachel(1984) and Hakistan, Rogers and Cattell(1982) for non-normally distributed variables. Another method described in Spirtes, Glymour, and Scheines(1993) and Scheines(1993) uses tests of vanishing tetrad differences among the measured variables. A tetrad difference among four measured variables is of the form $\rho(X_1,X_2)\times\rho(X_3,X_4)$ –

$\rho(X_1,X_4)\times\rho(X_2,X_3)$, where $\rho(X_1,X_2)$ represents the correlation of X_1 and X_2. For a given set of four variables, there are 3 different tetrad differences among the four variables (which can be formed by permuting the order of the variables in the example given). There is a one factor linear model of the four variables X_1, X_2, X_3, and X_4 if and only if all three tetrad differences among the four variables are equal to zero. There is a statistical test for vanishing tetrad differences for normally distributed variables devised by Wishart(1928), and an asymptotically correct distribution free test devised by Bollen(1990). A simultaneous test for multiple vanishing tetrad differences can be approximated by making a Bonferroni adjustment (see Bollen 1990). The advantage of testing the tetrad differences is that it does not require maximum likelihood estimates of the parameters, which can be slow and suffer from convergence problems. By increasing the sample size and decreasing the significance level it is possible to reduce the probability of type I and type II errors simultaneously.

In the discrete case, whether the one factor places constraints on the measured marginal or not depends upon the number of categories for the measured and latent variables. Bartholomew(1980), Holland(1981), and Rosenbaum(1984) describe tests for one factor models assuming the measured variables are binary, and the probability of a positive response increases monotonically with increasing values of the latent variable. A more general method of testing whether there is a one factor discrete model is to compare the observed frequencies with those predicted by the model using standard χ^2 goodness of fit tests, as in Goodman(1978). (Each one factor model, in addition to being a discrete DAG model, is also a graphical log-linear model. See Bishop, Fienberg, and Holland, 1975 and Whittaker, 1990 for descriptions of graphical log-linear models.) Unfortunately, as Bartholomew(1985) points out, this method is not practical when the number of variables is large. Aitkin et. al. (1981) used a graphical method to test goodness of fit, but there is no completely satisfactory method when the number of variables is large.

5.4 Correctness of the Algorithm

Three different kinds of problems can arise when the purification algorithm is applied. First, the algorithm might remove so many variables from **M'** that given the distributional assumptions (linear or discrete) the only one factor models that can be formed from the variables remaining in **M'** are not overidentified, and cannot be subjected to a statistical test. In such a case, there can obviously be no guarantee that the true measurement DAG for **M'** is pure. In the linear case, as long as there are at least two latent variables, and each latent variable has at least three indicator in **M'**, this is not a problem. In the discrete case, whether or not there are overidentified one factor models of $n \times 1$ subsets of measured variables in **M'** depends upon the number of categories of the measured and latent variables.

Second, it might be that the true measurement DAG does not entail that there exist one factor models for each $n \times 1$ subset, but it happens to have a parameterization in which each $n \times 1$ subset has a one factor model of the appropriate kind (linear or discrete). In Spirtes, Glymour and Scheines(1993) it is shown that the set of linear parameterizations of G for which tetrad differences are equal to zero when they are not entailed by G to be equal to 0, has Lebesgue measure 0. I do not know of an analogous result for the disctete case.

Third, it is possible that the true measurement DAG entails the existence of the appropriate one factor models even though it is not pure. However, it turns out that if the true measurement DAG entails the existence of the appropriate one factor models, it is guaranteed to be almost pure in the sense defined below, and almost purity is a sufficient condition for the tests of conditional independence described in the next section to be correct.

A DAG G is **almost pure** with respect to function *Latent* and a partition of **V** into **M**, **L**, and **O** if and only if:

1. if M is in **Measured**(L) then L is a direct cause (parent in the causal graph) of M with respect to $\mathbf{M} \cup \mathbf{L} \cup \mathbf{O}$, and
2. for every L in **L**, **Measured**(L) is not empty, and
3. for each $M \in \mathbf{M}$, G entails that M is independent of $(\mathbf{M} \cup \mathbf{L})\backslash\{M, Latent(M)\}$ given $Latent(M)$.

For the case where the probabilty measure over the measured variables is the marginal of a probability measured generated by a linear RSEM, the next theorem states necessary and sufficient conditions for a causal DAG to be almost pure. If **V** has a correlation matrix C, and $\mathbf{S} \subseteq \mathbf{V}$, let $C(\mathbf{S})$ be the marginal correlation matrix of C.

Theorem 1: If G is a DAG over a set of variables $\mathbf{M} \cup \mathbf{L} \cup \mathbf{O}$, **L** contains more than one variable, and $C(\mathbf{M} \cup \mathbf{L} \cup \mathbf{O})$ is generated by a linear parameterization of G, then G contains an almost pure measurement DAG with respect to function *Latent* and a partition of **V** into **M**, **L**, and **O** if and only if

1. if M is in **Measured**(L) then L is a direct cause (parent in the causal graph) of M with respect to $\mathbf{M} \cup \mathbf{L} \cup \mathbf{O}$;
2. for each $L \in \mathbf{L}$, there are at least 3 measured variables in **Measured**(L);
3. for every linear parameterization of G and each 3×1 foursome $\mathbf{S}$ of measured variables there is a linear parameterization of a one factor DAG of $\mathbf{S}$ with $C(\mathbf{S})$.

This entails that given a prior distribution over DAGs in which the true causal DAG has a non-zero probability, and for each DAG G the distribution over the linear parameters not fixed at zero and the variances of the error terms is absolutely continuous with Lebesgue measure, the probability is one that a set of measured variables is almost pure given that all of the $n \times 1$ subsets have one factor line models.

A weaker result is available for the discrete case. Theorem 2 states a necessary condition for a DAG to be almost pure.

Theorem 2: If G is a DAG over a set of variables $\mathbf{M} \cup \mathbf{L} \cup \mathbf{O}$, $P(\mathbf{M} \cup \mathbf{L} \cup \mathbf{O})$ is generated by a discrete parameterization of G, and G is an almost pure measurement DAG with respect to function *Latent* and a partition of $\mathbf{V}$ into $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{O}$ then for every discrete parameterization of G, for each $L \in \mathbf{L}$, and for each subset $\mathbf{S} \subseteq$ **Measured**(L), and single variable $M \in \mathbf{M} \backslash \mathbf{S}$ there is a discrete parameterization of a one factor DAG of $\mathbf{S} \cup \{M\}$ with marginal probability measure $P(\mathbf{S} \cup \{M\})$, where the latent variable has the same number of categories as L.

At the end of the purification process for discrete variables, one can test whether the measurement DAG is almost pure by testing whether there is a parameterization of a DAG with a pure measurement DAG and all of the latent variables adjacent to each other that fits the data. If there is no such parameterization, conclude that the subset of measured variables does not have an almost pure measurement DAG. However, at that point more heuristics are needed for eliminating further variables to find a subset of measured variables that do have an almost pure measurement DAG.

6. Testing For Conditional Independence Among Latent Variables

In this section, the inputs, outputs, and basic structure of an algorithm for tesing conditional independence among latent variables is described.

6.1 The Input

The input to the algorithm is:

1. sample data for a set of measured variables **M'**;
2. an almost pure measurement DAG for **M'**;
3. an assumption about the family of distributions that the true model lies in, either discrete (including a specification of the maximum number of categories for each latent variable), or linear;
4. a specification of the conditional independence relation to be tested;
5. a significance level for the test to be performed.

6.2 The Output of the Algorithm

The output of the algorithm is a yes or no decision about whether the specified conditional independence relation is judged to hold in the population.

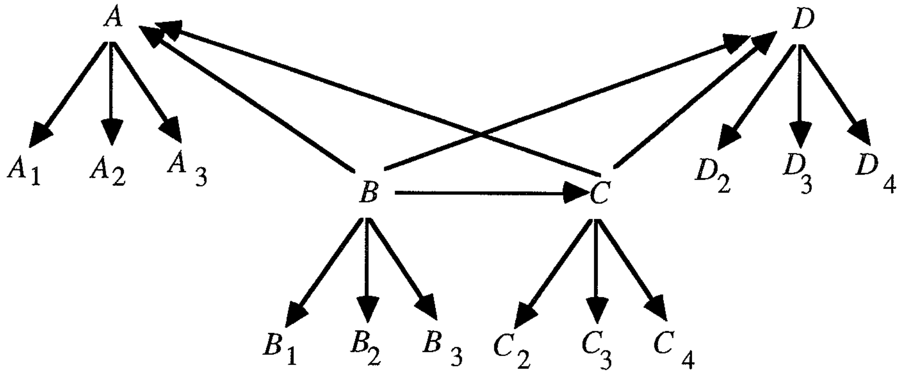
6.3 Outline of the Algorithm

Given an almost pure DAG G and a probability measure generated by G with density function $f(\mathbf{V})$, in order to test whether two latent variables A and B are independent conditional on a set of latent variables $\mathbf{Q}$, a DAG $Test(G,A,B,\mathbf{Q})$ will be formed. This DAG entails exactly one non-trivial[4] conditional independence relation among the latent variables, namely that A and B are independent conditional on $\mathbf{Q}$. $Test(G,A,B,\mathbf{Q})$ is constructed in such a way that A is independent of B given $\mathbf{Q}$ in $f(\mathbf{V})$ if and only if there is a parameterization of $Test(G,A,B,\mathbf{Q})$ such that G has the same marginal as $f(\mathbf{V})$ over the variables in $Test(G,A,B,\mathbf{Q})$.

If DAG G contains a set of variables $\mathbf{V}$, and G is an almost pure DAG with respect to function *Latent* and a partition of $\mathbf{V}$ into $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{O}$, let $\mathbf{V}(A,B,\mathbf{Q}) = \{A,B\} \cup \mathbf{Q} \cup$ **Measured**$(\{A,B\} \cup \mathbf{Q})$. If A and B are in $\mathbf{L}$ and $\mathbf{Q}$ is included in $\mathbf{L}$, then the set of vertices in $Test(G,A,B,\mathbf{Q})$ is the set $\mathbf{V}(A,B,\mathbf{Q})$; each pair of variables in $\mathbf{Q}$ is adjacent; for each L in $\{A,B\} \cup \mathbf{Q}$ and each member M of **Measured**(L), there is an edge from L to M; there are edges from each member of $\mathbf{Q}$ to A and B; there is no edge between A and B. For example, if G has the measurement DAG shown in figure 3, then $Test(G,A,D,\{B,C\})$ is shown in figure 5. $Test'(G,A,B,\mathbf{Q})$ is the same DAG as $Test(G,A,B,\mathbf{Q})$ except that $Test'(G,A,B,\mathbf{Q})$ also contains an edge from A to B. It is assumed that $Test'(G,A,B,\mathbf{Q})$ is

[4] A conditional independence relation among variables in $\mathbf{V}$ is non-trivial if is not true in every probability distribution over $\mathbf{V}$.

identified, which in the linear case is always true as long as each latent variable contains at least three measured indicators.



Test (*G* ,*A* ,*D*, {*B* ,*C* })

Figure 5

Let $\rho(A,B.\mathbf{Q})$ represent the partial correlation of A and B given $\mathbf{Q}$. The algorithm for testing whether $\rho(A,B.\mathbf{Q}) = 0$ in the linear case, or A and B are independent conditional on $\mathbf{Q}$ in the discrete case is basically the same: if *Test*'(G,A,B,$\mathbf{Q}$) is an identified model, and after consistently estimating the parameters in *Test*'(G,A,B,$\mathbf{Q}$) and *Test*(G,A,B,$\mathbf{Q}$), the fit of the models to the data is significantly different, the corresponding constraint is judged not to hold.

In the linear case, there are several different ways of testing if the difference in fit is significant. One could determine whether the difference in fit of the the two nested models is significant, using programs such as LISREL (Joreskog and Sorbom, 1984) or EQS (Bentler, 1985). (There are such tests for normally distributed variables, as well as asymptotically distribution free tests.) In the linear normal case, one could perform a maximum likelihood estimate of the parameters of *Test*(G,A,B,$\mathbf{Q}$), and perform a t-test to see if the parameter of the edge between A and B is significantly different from zero.

If there are at least three measured indicators for each latent variable, then *Test*'(G,A,B,$\mathbf{Q}$) is an identified linear model, and *Test*(G,A,B,$\mathbf{Q}$) is overidentified. Scheines (personal communication) has pointed out that in testing for zero partial correlations among the latent variables it is always possible to estimate the correlation matrix among the latents in *Test*'(G,A,B,$\mathbf{Q}$) as long as it is identified. Hence, an alternative procedure is to test for zero partial correlations among the latents by estimating the correlation matrix among the latents, and then directly testing for zero partial correlations. The test should take into account that the correlation matrix among the latent variables was estimated rather than measured. Which of these two procedures is more reliable in small samples is not known.

In the discrete case one could compare *Test*(*G*,*A*,*B*,**Q**) with *Test*(*G*,*A*,*B*,**Q**) by determining whether the difference in the fit of the two nested models is significant. The parameters can be estimated using the E-M algorithm described in Bartholomew(1985). Unfortunately, as Bartholomew(1985) points out, the standard tests of goodness of fit are not practical when the number of variables is large.

6.4 Correctness of the Algorithm

The tests of conditional independence or zero partial correlation have been reduced to tests of discrete or linear models. For these tests of models, in the normal case, the calculation of the probability of type I error is correct, and in the discrete case and the non-normal linear case, the calculation of the probability of type I error is asymptotically correct. The correctness of the calculation of the probability of type I error for the tests of zero partial correlation or conditional independence is guaranteed by the following two theorems.

Theorem 3: If *G* is an almost pure DAG with respect to function *Latent* and a partition of **V** into **M**, **L**, and **O**, *G* has a linear parameterization with marginal correlation matrix *C*(**Measured**({*A*,*B*} ∪ **Q**}), and *Test'*(*G*,*A*,*B*,**Q**) is identified, then $\rho(A,B.\mathbf{Q}) = 0$ in *C* if and only there is a linear parameterization of *Test*(*G*,*A*,*B*,**Q**) with marginal correlation matrix *C*(**Measured**({*A*,*B*} ∪ **Q**}).

Adapting the notation in Pearl(1988) write $I_P(\mathbf{X},\mathbf{Y},\mathbf{Z})$ if and only if **X** is independent of **Z** conditional on **Y** in a probability measure *P*

Theorem 4: If *G* is an almost pure DAG with respect to function *Latent* and a partition of **V** into **M**, **L**, and **O**, *G* has a discrete parameterization with marginal *P*(**Measured**({*A*,*B*} ∪ **Q**)), and *Test'*(*G*,*A*,*B*,**Q**) is identified, then $I_P(A,\mathbf{Q},B)$ if and only there is a discrete parameterization of *Test*(*G*,*A*,*B*,**Q**) with marginal *P*(**Measured**({*A*,*B*} ∪ **Q**)).

If the variables in **M** ∪ **L** are discrete, then the theorem holds when both *Test'(G*,*A*,*B*,**Q**) and *Test(G*,*A*,*B*,**Q**) are also discrete DAG models, with the latent variables in *Test'(G*,*A*,*B*,**Q**) and *Test(G*,*A*,*B*,**Q**) having the same number of categories as the corresponding variables in {*A*,*B*} ∪ **Q**. (*Test'*(*G*,*A*,*B*,**Q**) and *Test*(*G*,*A*,*B*,**Q**), in addition to being discrete DAG models, are also graphical log-linear models.) Whether $\mathit{Test}'(G,A,B,\mathbf{Q})$ is identified depends upon the number of categories of the measured and latent variables.

7. Simulation Tests

In this simulation I tested only the PC algorithm using the test of zero partial correlation for the linear case described in section 6. For simulation tests of the purification algorithm, see Scheines, Spirtes, and Glymour (forthcoming).

At sample size 250, ten linear models were generated pseudo-randomly by the following process. Ten DAGs were generated, where each DAG had a pure measurement DAG in which there were six latent vertices and each latent vertex had four measured indicators. The adjacencies between the latent variables were pseudo-randomly generated, where on average each latent vertex was adjacent to 2 other latent vertices (i.e. the average order of vertices in the latent subgraph was 2). Each of the error terms was given a standard normal distribution. The linear coefficients for the existing edges were pseudo-randomly assigned values between 0.5 and 1.5. Then a pseudo-random sample was generated from each of the ten models. The sample data was input to the algorithm. Tests of conditional independence were done with the LISREL program. For each DAG G, and each partial correlation $\rho(A,B,\mathbf{Q}) = 0$ tested, the model $\mathit{Test}(G,A,B,\mathbf{Q})$ was formed, and given to LISREL along with the sample data. The parameters were estimated, and if LISREL judged (using its modification indices) that adding the edge between A and B would significantly improve the fit of the data the partial correlation was judged to be non-zero. The pattern output by the algorithm (the "output pattern") was then compared with the pattern corresponding to the DAG that generated the data (the "true pattern"). Four kinds of errors were counted. If latent variables A and B were adjacent in the output pattern but not in the true pattern, this was counted as an edge error of commission. If latent variables A and B were adjacent in the true pattern but not in the output pattern, this was counted as an edge error of omission. If A and B were adjacent in both the true pattern and the output pattern, but the edge between A and B had an arrowhead at one end in the output pattern but not the true pattern, this was counted as an arrowhead error of commission. Finally, if A and B were adjacent in both the true pattern and the output pattern, but the edge between A and B had an arrowhead at one end in the true pattern but not the output pattern, this was counted as an arrowhead error of omission. This process was then repeated at sample sizes of 1000, 2500, and 5000, and also with the average order of vertices in the latent subgraph set to 3. The results are summarized in the following tables.

		Average Order	
		2	3
Sample Size	250	35.1	41.3
	1000	17.4	28.9
	2000	12.8	20.4

Table 1: % Edge Errors of Commission

		Average Order	
		2	3
Sample Size	250	6.4	4.6
	1000	1.2	3.7
	2000	4.2	7.5

Table 2: % Edge Errors of Omission

		Average Order	
		2	3
Sample Size	250	33.1	33.3
	1000	11.1	16.1
	2000	9.3	17.7

Table 3: % Arrow Errors of Commission

		Average Order	
		2	3
Sample Size	250	10.4	28.9
	1000	7.2	20.2
	2000	12.1	28.9

Table 4: % Arrow Errors of Omission

In general, the algorithm is more accurate on adjacencies than orientations. The overall performance tends to improve with increasing sample size. The percentage of arrow errors of omission and commission is quite high even at fairly large sample sizes, when the graph is not sparse.

8. Conclusion

The strategy described here for detecting causal relations among latent variables is in some respects quite general. With respect to the purification algorithm, if G is an almost pure DAG model, then for any latents L_1 and L_2, there are one factor models of all subsets of measured variables that contain a subset of the indicators of L_1 and a single indicator of L_2. However, the mere existence of a one factor model does not in general place any constraints on the observed data, unless the family of probability measures that the one factor model lies in is restricted. This is most plausibly done by relating the family of probability measures of the one factor model to the family of probability measures the model generated by DAG G is assumed to lie in. This is what has been done for the linear and discrete cases. An open question is whether the same kind of relationship between the one factor models and the family of probability measures the model generated by G is assumed to lie in can be demonstrated for other interesting families of probability measures. Similarly, it is an open question whether the relationship between $Test(G,A,B,\mathbf{Q})$ or $Test'(G,A,B,\mathbf{Q})$ and the family of probability measures the model generated by G is assumed to lie in can be demonstrated for other interesting families of probability measures.

References

Aitkin, M., Anderson, D. and Hinde. J. (1981). "Statistical modelling of data on teaching styles", *J. Roy. Statist. Soc.,* A, 144, pp. 419-461.

Akaike, H. (1983). "Information measure and model selection", *Bull. Int. Statist. Inst.,* 50, Book I, pp. 277-90.

Amemiya, Y., and Anderson, T. (1985). "Asymptotic chi-square tests for a large class of factor analysis models", Technical Report No. 13, Econometric Workshop, Stanford Univ.

Anderson, J., and Gerbing, D. (1982). "Some methods for respecifying measurement models to obtain unidimensional construct measurement", *Journal of Marketing Research.*, 19, pp. 453-60.

Anderson, J., and Gerbing, D. (1988). "Structural equation modeling in practice: A review and recommended two-step approach," *Psychological Bulletin,* 103, pp. 411-423.

Bartholomew, D. (1987). *Latent Variable Models and Factor Analysis,* Oxford University Press, NY.

Bartholomew, D. (1980). "Factor analysis on categorical data", *J. Roy. Statist. Soc., B,* 42, pp. 93-99.

Bentler, P. (1985). *Theory and Implementation of EQS: A Structural Equations Program.*, BMDP Statistical Software Inc., Los Angeles.

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

Bollen, K. (1990). "Outlier screening and a distribution-free test for vanishing tetrads", *Sociological Methods and Research* 19, pp. 80-92.

Bollen, K. (1989). *Structural Equations with Latent Variables.* Wiley, NY.

Bozdogan, H. and Ramirez, D. (1986). *Model selection approach to the factor model problem, parameter parsimony, and choosing the number of factors.* Research Report, Dept. of Mathematics, Univ. of Virginia, Charlottesville, VA.

Fachel, J. (1986). *The C-type Distribution as an Underlying Model for Categorical Data and its use in Factor Analysis,* Ph.D. Thesis, University of London.

Goodman, L. (1978) *Analyzing Qualitative/Categorical Data,* (ed. by J. Magidson), ABT Books, Cambridge, MA.

Hakstian, A., Rogers, W., and Cattell, R. (1982). "The behavior of number-of-factors rules with simulated data", *Multivariate Behavioral Research,* 17, pp. 193-219.

Holland, P. (1981). "When are item response models consistent with observed data?", *Psychometrika,* 46, pp. 79-92.

Joreskog, K. and Sorbom, D. (1984). *LISREL VI User's Guide*. Scientific Software, Inc., Mooresville, IN.

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

Kiiveri, H., Speed, T., and Carlin, J. (1984). "Recursive causal models." *Journal of the Australian Mathematical Society* **36**, 30-52.

Lauritzen, S., Dawid, A., Larsen, B., Leimer, H. (1990). "Independence Properties of Directed Markov Fields." *Networks*, **20**, 491-505.

Muthen, B. (1978). "Contributions to factor analysis of dichotomous variables", *Psychometrika*, 43, pp. 551-60.

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

Rosenbaum, P. (1984). "Testing the conditional independence and monotonicity assumptions of item response theory", *Psychometrika*, 49, pp. 425-435.

Scheines, R. (1993). *Unidimensional Linear Latent Variable Models,* Technical Report CMU-PHIL-39, Department of Philosophy, Carnegie Mellon University, Pgh, PA.

Scheines, R., Spirtes, P., Meek, C., and Glymour, C., (forthcoming). *TETRAD II: Tools for Causal Modelling*, Lawrence Erlbaum, NY.

Schwarz, G. (1978). "Estimating the dimension of a model", *Ann. Statist.*, 6, pp. 461-4.

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

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

Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *JASA* 75, 963-972.

Wermuth, N. and Lauritzen, S. (1983). Graphical and recursive models for contingency tables. *Biometrika* 72, 537-552.

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

Wishart, J. (1928). "Sampling errors in the theory of two factors", *British Journal of Psychology* , 19, pp. 180-187.

Wright, S. (1934). "The method of path coefficients", *Ann. Math. Stat.* 5, pp. 161-215.

Appendix

There is a graphical relation, d-separation, which characterizes when a DAG entails that **X** is independent of **Y** given **Z**, which is used in the following proofs. If **X**, **Y**, and **Z** are disjoint sets of variables in a DAG G, then **X** and **Y** are **d-separated** given **Z** if and only if there is no undirected path U from any X in **X** to any Y in **Y** such that every collider on U has a descendant in **Z**, and no non-collider on U is in **Z**; otherwise say that **X** and **Y** are **d-connected** given **Z**. If **X**, **Y**, and **Z** are disjoint then DAG G entails **X** is independent of **Y** given **Z** if and only if **X** is d-separated from **Y** given **Z**. See Pearl (1988). In a linear DAG model, if $X \neq Y$, and X and Y are not in **Z** then a DAG G entails $\rho(X,Y.\mathbf{Z}) = 0$ for all linear parameterization of G in which $\rho(X,Y.\mathbf{Z})$ is defined and the errors are uncorrelated and have non-zero variance if and only if X is d-separated from Y given **Z**. See Spirtes, Glymour, and Scheines (1993).

The proofs below depend upon the following properties of linear models and discrete models. Given a DAG G with variables **V**, as long as G does not constrain some partial correlation to be zero when it is not zero in a correlation matrix C, there is a linear parameterization of G with correlation matrix C. (This parameterization can be formed by simply making the linear coefficient of B in the equation for A equal to the partial regression coefficient of A on B given $\mathbf{V}\backslash\{A,B\}$.) Similarly, given a DAG G with variables **V**, as long as G does entail some conditional independence relation false in probability measure $P(\mathbf{V})$, there is a discrete parameterization of G with probability measure $P(\mathbf{V})$.

A DAG H is an **I-map** of a density $f(\mathbf{V})$ if and only if it contains vertices **V** and for all disjoint **X**, **Y**, and $\mathbf{Z} \subseteq \mathbf{V}$, if **X** and **Y** are d-separated given **Z** then $I_f(\mathbf{X},\mathbf{Z},\mathbf{Y})$. If H is not an I-map of $f(\mathbf{V})$, then there is no parameterization of H with density $f(\mathbf{V})$ because H entails that **X** is independent of **Y** given **Z**, but $\sim I_f(\mathbf{X},\mathbf{Z},\mathbf{Y})$.

Lemma 1: If G is an almost pure measurement DAG with respect to function *Latent* and a partition of **V** into **M**, **L**, and **O**, and G has a parameterization with density $f(\mathbf{V})$, then G entails that for each $L \in \mathbf{L}$, and each subset $\mathbf{S} \subseteq$ **Measured**(L), and single variable $M \in \mathbf{M}\backslash\mathbf{S}$ a one factor DAG of $\mathbf{S} \cup \{M\}$ is an I-map of $P(\mathbf{S} \cup \{M\})$.
Proof. G entails that each variable M in $\mathbf{S} \subseteq$ **Measured**(L) is independent from all of the other variables in $(\mathbf{M} \cup \mathbf{L})\backslash\{M,L\}$ given L. Hence there is a factorization of $f(\mathbf{S} \cup \{M\})$ of the form:

$$f(\mathbf{V}) = f(\mathbf{S} \cup \{M,L\}) \times f(\mathbf{V} \setminus (\mathbf{S} \cup \{M,L\}) \mid \mathbf{S} \cup \{M,L\})$$

It is possible to calculate $f(\mathbf{S} \cup \{M,L\})$ by integrating out $\mathbf{V} \setminus (\mathbf{S} \cup \{M,L\})$. Because each variable in $f(\mathbf{S} \cup \{M\})$ is independent of the other variables given L,

$$f(\mathbf{S} \cup \{M,L\}) = \prod_{X \in \mathbf{S} \cup \{M\}} f(X \mid L) \times f(L)$$

This is a parameterization of a one factor model with latent variable L. Hence the one factor DAG is an I-map of $f(\mathbf{S} \cup \{M,L\})$. Q.E.D

Theorem 2: If G is a DAG over a set of variables $\mathbf{M} \cup \mathbf{L} \cup \mathbf{O}$, $P(\mathbf{M} \cup \mathbf{L} \cup \mathbf{O})$ is generated by a discrete parameterization of G, and G is an almost pure measurement DAG with respect to function *Latent* and a partition of $\mathbf{V}$ into $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{O}$ then for every discrete parameterization of G, for each $L \in \mathbf{L}$, and for each subset $\mathbf{S} \subseteq$ **Measured**(L), and single variable $M \in \mathbf{M} \setminus \mathbf{S}$ there is a discrete parameterization of a one factor DAG of $\mathbf{S} \cup \{M\}$ with marginal probability measure $P(\mathbf{S} \cup \{M\})$, where the latent variable has the same number of categories as L.

Proof. Suppose that G is almost pure. By lemma 1, the one factor model with latent L is an I-map of $P(\mathbf{S} \cup \{M,L\})$. Hence the one factor DAG does not entail any conditional independence relations false in $P(\mathbf{S} \cup \{M,L\})$. It follows that there is a discrete parameterization of the one factor DAG with probability measure $P(\mathbf{S} \cup \{M,L\})$. This probability measure has marginal probability measure $P(\mathbf{S} \cup \{M\})$. Q.E.D.

Lemma 2: If G is an almost pure model with respect to function *Latent* and a partition of $\mathbf{V}$ into $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{O}$, and G has a parameterization with marginal density $f(\{A,B\} \cup \mathbf{Q})$, then *Test'*$(G,A,B,\mathbf{Q})$ is an I-map of density $f(\mathbf{V}(A,B,\mathbf{Q}))$.

Proof. By definition of almost pure:

$$f(\mathbf{M} \cup \mathbf{L}) = \left(\prod_{M \in \mathbf{M}} f(M | Latent(M)) \right) \times f(\mathbf{L}) =$$

$$\left(\prod_{M \in \mathbf{Measured}(\{A,B\} \cup \mathbf{Q})} f(M | Latent(M) \right) \times f(\{A,B\} \cup \mathbf{Q}) \times$$

$$\left(\prod_{M \in \mathbf{Measured}(\mathbf{L} \setminus (\{A,B\} \cup \mathbf{Q}))} f(M | Latent(M)) \right) \times f(\mathbf{L} \setminus (\{A,B\} \cup \mathbf{Q}) | (\{A,B\} \cup \mathbf{Q}))$$

If the variables in $\mathbf{V} \setminus \mathbf{V}(A,B,\mathbf{Q})$ are integrated out then

$$f(\mathbf{V}(A,B,\mathbf{Q})) = \left(\prod_{M \in \mathbf{Measured}(\{A,B\} \cup \mathbf{Q}\}} f(M | Latent(M) \right) \times f(\{A,B\} \cup \mathbf{Q})$$

But this is also a parameterization of *Test'*$(G,A,B,\mathbf{Q})$. Hence *Test'*$(G,A,B,\mathbf{Q})$ does not entail any conditional independence relations false in $f(\mathbf{V}(A,B,\mathbf{Q}))$, and is an I-map of $f(\mathbf{V}(A,B,\mathbf{Q}))$. Q.E.D.

Lemma 3: If G is an almost pure model with respect to function *Latent* and a partition of $\mathbf{V}$ into $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{O}$, G has a parameterization with marginal density $f(\{A,B\} \cup \mathbf{Q})$, and $I_f(A,\mathbf{Q},B)$ then *Test*$(G,A,B,\mathbf{Q})$ is an I-map of $f(\mathbf{V}(A,B,\mathbf{Q}))$.

Proof. By lemma 2, there is a factorization of $f(\{A,B\} \cup \mathbf{Q})$ of the form

$$f(\mathbf{V}(A,B,\mathbf{Q})) = \left(\prod_{M \in \mathbf{Measured}(\{A,B\} \cup \mathbf{Q}\}} f(M | Latent(M) \right) \times f(\{A,B\} \cup \mathbf{Q})$$

Since $I_f(A,\mathbf{Q},B)$, it follows that

$$f(\mathbf{V}(A,B,\mathbf{Q})) = \left(\prod_{M \in \mathbf{Measured}(\{A,B\} \cup \mathbf{Q}\}} f(M | Latent(M) \right) \times f(A | \mathbf{Q}) \times f(B | \mathbf{Q}) \times f(\mathbf{Q})$$

But this is also a parameterization of *Test*$(G,A,B,\mathbf{Q})$. Hence *Test*$(G,A,B,\mathbf{Q})$ entails a subset of the conditional independence relations true in $f(\mathbf{V}(A,B,\mathbf{Q}))$, and is an I-map of $f(\mathbf{V}(A,B,\mathbf{Q}))$. Q.E.D.

Theorem 3: If G is an almost pure DAG with respect to function *Latent* and a partition of $\mathbf{V}$ into $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{O}$, G has a linear parameterization with marginal correlation matrix $C(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q})$, and $\mathit{Test}'(G,A,B,\mathbf{Q})$ is identified, then $\rho(A,B.\mathbf{Q}) = 0$ in C if and only there is a linear parameterization of $\mathit{Test}(G,A,B,\mathbf{Q})$ with marginal correlation matrix $C(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q}))$.

Proof. Suppose that the probability measure generated by G has density $f(\mathbf{V})$ with correlation matrix C, and $\rho(A,B.\mathbf{Q}) = 0$ in C. C depends only upon the variances of the exogenous variables and the linear coefficients, so there is a density f' which is jointly normal and has the same correlation matrix C. In f', $\rho(A,B.\mathbf{Q}) = 0$, so $I_f(A,\mathbf{Q},B)$. By lemma 3, $\mathit{Test}(G,A,B,\mathbf{Q})$ is an I-map of $f'(\mathbf{V}(A,B,\mathbf{Q}))$. Hence $\mathit{Test}(G,A,B,\mathbf{Q})$ does not entail any conditional independence relations false in $f'(\mathbf{V}(A,B,\mathbf{Q}))$. It follows that $\mathit{Test}(G,A,B,\mathbf{Q})$ does not constrain any partial correlations to be zero when they are not zero in $C(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q}\})$. Hence there is a linear parameterization of $\mathit{Test}(G,A,B,\mathbf{Q})$ with marginal correlation matrix $C(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q}))$.

Suppose that the probability measure generated by G has density $f(\mathbf{V})$ with correlation matrix C, and $\rho(A,B.\mathbf{Q}) \neq 0$ in C. By lemma 2, $\mathit{Test}'(G,A,B,\mathbf{Q})$ is an I-map of $f(\mathbf{V})$ so there is a linear parameterization of $\mathit{Test}'(G,A,B,\mathbf{Q})$ with correlation matrix $C(\mathbf{V}(A,B,\mathbf{Q}))$, and because it is identified the parameterization is unique. In that parameterization, $\rho(A,B.\mathbf{Q}) \neq 0$. Suppose that there is a parameterization of $\mathit{Test}(G,A,B,\mathbf{Q})$ with marginal correlation matrix $C(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q}))$. In every parameterization of $\mathit{Test}(G,A,B,\mathbf{Q})$, $\rho(A,B.\mathbf{Q}) = 0$. Because $\mathit{Test}'(G,A,B,\mathbf{Q})$ and $\mathit{Test}(G,A,B,\mathbf{Q})$ are nested models, it follows that there is a parameterization of $\mathit{Test}'(G,A,B,\mathbf{Q})$ with marginal correlation matrix $C(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q}\})$ in which $\rho(A,B.\mathbf{Q}) = 0$. This is a contradiction. Q.E.D.

Theorem 4: If G is an almost pure DAG with respect to function *Latent* and a partition of $\mathbf{V}$ into $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{O}$, G has a discrete parameterization with marginal $P(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q}))$, and $\mathit{Test}'(G,A,B,\mathbf{Q})$ is identified, then $I_P(A,\mathbf{Q},B)$ if and only there is a discrete parameterization of $\mathit{Test}(G,A,B,\mathbf{Q})$ with marginal $P(\mathbf{Measured}(\{A,B\} \cup \mathbf{Q}))$.

Proof. Suppose that $I_P(A,\mathbf{Q},B)$. By lemma 3, $\mathit{Test}(G,A,B,\mathbf{Q})$ is an I-map of $P(\mathbf{V}(A,B,\mathbf{Q}))$. Hence it does not entail any conditional independence relations false in $P(\mathbf{V}(A,B,\mathbf{Q}))$. It follows that there is a discrete parameterization of $\mathit{Test}(G,A,B,\mathbf{Q})$ with probability measure $P(\mathbf{V}(A,B,\mathbf{Q}))$.

Suppose that $\sim I_P(A,\mathbf{Q},B)$. By lemma 2, $\mathit{Test}'(G,A,B,\mathbf{Q})$ is an I-map of $P(\mathbf{V}(A,B,\mathbf{Q}))$. Hence there is a discrete parameterization of $\mathit{Test}'(G,A,B,\mathbf{Q})$ with probability measure

P(**V**(*A*,*B*,**Q**)). Because *Test'*(*G*,*A*,*B*,**Q**) is identified, the parameterization is unique. Suppose now that there is a parameterization of *Test*(*G*,*A*,*B*,**Q**) with marginal probability measure *P*(**Measured**({*A*,*B*} $\cup$ **Q**)). In that parameterization *A* is independent of *B* given **Q**. Because *Test*(*G*,*A*,*B*,**Q**) is a proper subgraph of *Test'*(*G*,*A*,*B*,**Q**), that is also a parameterization of *Test'*(*G*,*A*,*B*,**Q**) in which *A* is independent of *B* given **Q**. But that is a contradiction. Q.E.D.