27

Automated Search for Causal Relations - Theory and Practice

PETER SPIRTES, CLARK GLYMOUR, RICHARD SCHEINES, AND ROBERT TILLMAN

1 Introduction

The rapid spread of interest in the last two decades in principled methods of search or estimation of causal relations has been driven in part by technological developments, especially the changing nature of modern data collection and storage techniques, and the increases in the speed and storage capacities of computers. Statistics books from 30 years ago often presented examples with fewer than 10 variables, in domains where some background knowledge was plausible. In contrast, in new domains, such as climate research where satellite data now provide daily quantities of data unthinkable a few decades ago, fMRI brain imaging, and microarray measurements of gene expression, the number of variables can range into the tens of thousands, and there is often limited background knowledge to reduce the space of alternative causal hypotheses. In such domains, non-automated causal discovery techniques appear to be hopeless, while the availability of faster computers with larger memories and disc space allow for the practical implementation of computationally intensive automated search algorithms over large search spaces. Contemporary science is not your grandfather's science, or Karl Popper's.

Causal inference without experimental controls has long seemed as if it must somehow be capable of being cast as a kind of statistical inference involving estimators with some kind of convergence and accuracy properties under some kind of assumptions. Until recently, the statistical literature said *not.* While parameter estimation and experimental design for the effective use of data developed throughout the 20th century, as recently as 20 years ago the methodology of causal inference without experimental controls remained relatively primitive. Besides a cessation of hostilities from the majority of the statistical and philosophical communities (which has still only partially happened), several things were needed for theories of causal estimation to appear and to flower: well defined mathematical objects to represent causal relations; well defined connections between aspects of these objects and sample data; and a way to compute those connections. A sequence of studies beginning with Dempster's work on the factorization of probability distributions [Dempster 1972] and culminating with Kiiveri and Speed's [Kiiveri & Speed 1982] study of linear structural equation models, provided the first, in the form of directed acyclic graphs, and the second, in the form of the "local" Markov condition. Pearl and his students [Pearl 1988], and independently, Stefan Lauritzen and his collaborators [Lauritzen, Dawid, Larsen, & Leimer 1990], provided the

third, in the form of the "global" Markov condition, or d-separation in Pearl's formulation, and the assumption of its converse, which came to be known as "stability" or "faithfulness." Further fundamental conceptual and computational tools were needed, many of them provided by Pearl and his associates; for example, the characterization and representation of Markov equivalence classes and the idea of "inducing paths," essential to understanding the properties of models with unrecorded variables. Initially, most of these authors, including Pearl, rejected a causal interpretation of directed graphical models, or Bayes nets, in Pearl's case because of doubts about the possibility of reliable search when associations of measured variables might be due to unobserved confounders. Yet with the pieces Speed, Lauritzen, Pearl and others had established, a principled theory of causal estimation could, and did, begin around 1990, and Pearl and his students have made important contributions to it, (We did some nudging!). Pearl has become the foremost advocate in the universe for reconceiving the relations between causality and statistics. Once begun for special cases, the understanding of search methods for causal relations has expanded to a variety of scientific and statistical settings, and in many scientific enterprises—neuroimaging for example—causal representations and search are treated as almost routine.

Besides the development of estimation or search algorithms, and proofs of their properties, a theory of search for causal explanations required a theory of interventions that would both justify the causal interpretation of directed graphical models and also provide a coherent normative theory of inference using causal premises. That effort can be traced back to Strotz and Wold [Strotz & Wold 1960], then to our own work [Spirtes, Glymour, & Scheines 1993] on prediction from classes of causal graphs, and then to the full development of a non-parametric theory of prediction for graphical models by Pearl and his collaborators [Shpitser & Pearl 2008]. Pearl brilliantly turned philosopher and developed the theory of interventions into a general account of counterfactual reasoning. Although we will not discuss it further, we think there remain interesting open problems about prediction algorithms for various parametric classes of graphical causal models.

The following paper surveys a broad range of causal estimation problems and algorithms, concentrating especially on those that can be illustrated with empirical examples that we and our students and collaborators have analyzed. This has naturally led to a concentration on the algorithms and tools that we have developed. The kinds of causal estimation problems and algorithms discussed are broadly representative of the most important developments in methods for estimating causal structure since 1990, but it is not a comprehensive survey. There have been so many improvements to the basic algorithms that we describe here there is not room to discuss them all. A good resource for a description of further research in this area is the Proceedings of the Conferences on Uncertainty in Artificial Intelligence, at http://uai.sis.pitt.edu.

The dimensions of the problems, as we have long understood them, are these:

1. Finding computationally and statistically feasible methods for discovering causal information for large numbers of variables, provably correct under standard sampling assumptions, assuming no confounding by unrecorded variables.

2. The same when the "no confounding" assumption is abandoned.
3. Finding methods for obtaining causal information when there is systematic sample selection bias—when values of some of the variables of interest are associated with sample membership.
4. Finding methods for establishing the existence of unobserved causes and estimating *their* causal relations with one another.
5. Finding methods for discovering causal relations in data produced by feedback systems.
6. Finding methods for discovering causal relations in time series data.
7. Finding methods for discovering causal relations in linear and in non-linear non-Gaussian systems with continuous variables.
8. Finding methods for discovering causal relations using distributed, multiple data sets.
9. Finding methods for merging the above with experimental design.

2 Assumptions

We assume the reader's familiarity with the standard notions used in discussions of graphical causal model search: conditional independence, Markov properties, d-separation, Markov equivalence, patterns, distribution equivalence, causal sufficiency, etc. The appendix gives a brief review of the essential definitions, assumptions and theorems required for known proofs of correctness of the algorithms we will discuss.

3 Model Search Assuming Causal Sufficiency

The assumption of causal sufficiency (roughly no unrecorded confounders) is often unrealistic, but it is useful in explicating search because the concepts and methods used in search algorithms that make more realistic assumptions are more complex versions of ideas that are used in searches that assume causal sufficiency.

3.1 The PC Algorithm

The PC algorithm is a constraint-based search that attempts to find the pattern that most closely entails all and only the conditional independence constraints judged to hold in the population. The SGS algorithm [Spirtes & Glymour 1991] and the IC algorithm [Verma & Pearl 1990] were early versions of this algorithm which were statistically and computationally feasible only on data sets with few variables because they required conditioning on all possible subsets of variables.) The PC algorithm solved both difficulties in typical cases.

The PC algorithm has an adjacency phase in which the adjacencies are determined, and an orientation phase in which as many edges as possible are oriented. The adjacency phase is stated below, and illustrated in Figure 1. Let **Adjacencies**(G,A) be the set of vertices adjacent to A in undirected graph G. (In the algorithm, the graph G is continually updated, so **Adjacencies**(G,A) is constantly changing as the algorithm progresses.)

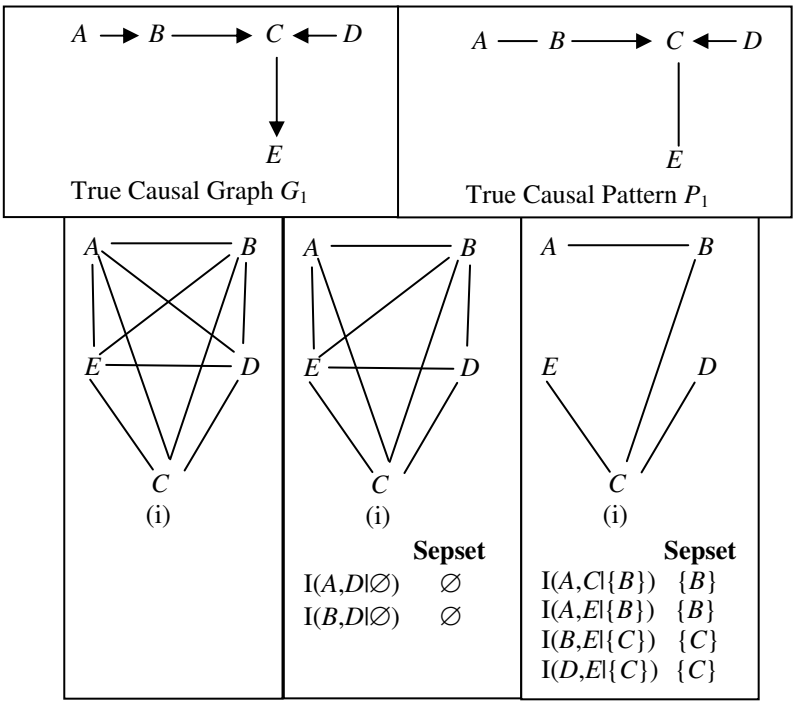


Figure 1: Constraint based search, where correct pattern is P_1

Adjacency Phase of PC Algorithm:
Form an undirected graph G in which every pair of vertices in **V** is adjacent.
$n := 0$.
repeat
  repeat
    Select an ordered pair of variables X and Y that are adjacent in G such that **Adjacencies**(G,X)\{Y} has cardinality greater than or equal to n, and a subset **S** of **Adjacencies**(G,X)\{Y} of cardinality n, and if X and Y are independent conditional on **S** delete edge X — Y from C and record **S** in **Sepset**(X,Y) and **Sepset**(Y,X);
  until all ordered pairs of adjacent variables X and Y such that **Adjacencies**(G,X)\{Y} has cardinality greater than or equal to n and all subsets **S** of **Adjacencies**(G,X)\{Y} of cardinality n have been tested for conditional independence;
  $n := n + 1$;
until for each ordered pair of adjacent vertices X, Y, **Adjacencies**(G,X)\{Y} is of cardinality less than n.

After the adjacency phase of the algorithm, the orientation phase of the algorithm is performed. The orientation phase of the algorithm is illustrated in

Figure 2.

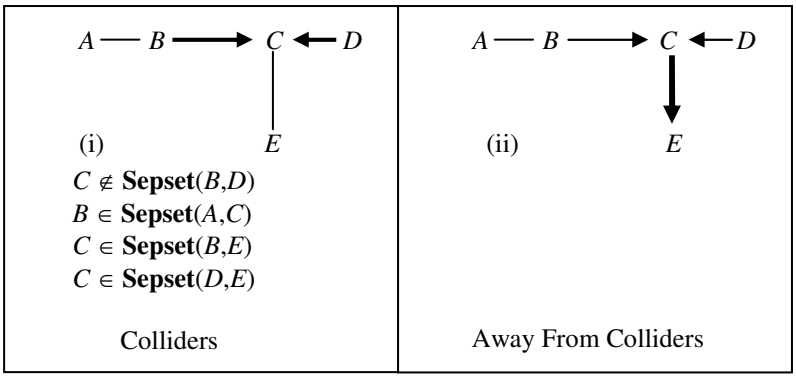


Figure 2: Orientation phase of PC algorithm, assuming true pattern is P_1

The orientation phase of the PC algorithm is stated more formally below. The last two orientation rules (Away from Cycles, and Double Triangle) are not used in the example, but are sound because if the edges were oriented in ways that violated the rules, there would be a directed cycle in the pattern, which would imply a directed cycle in the graph (which in this section is assumed to be impossible). The orientation rules are complete [Meek 1995], i.e. every edge that has the same orientation in every member of a DAG conditional independence equivalence class is oriented by these rules.

Orientation Phase of PC Algorithm

For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in graph G but the pair X, Z are not adjacent in G, orient $X — Y — Z$ as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in **Sepset**(X,Z).

repeat

Away from colliders: If $A \rightarrow B — C$, and A and C are not adjacent, then orient as $B \rightarrow C$.

Away from cycles: If $A \rightarrow B \rightarrow C$ and $A — C$, then orient as $A \rightarrow C$.

Double Triangle: If $A \rightarrow B \leftarrow C$, A and C are not adjacent, $A — D — C$, and there is an edge $B — D$, orient $B — D$ as $D \rightarrow B$.

until no more edges can be oriented.

The tests of conditional independence can be performed in the usual way. Conditional independence among discrete variables can be tested using the G^2 statistic; conditional independence among multivariate Gaussian variables can be tested using Fisher's z-transformation of the partial correlations [Spirtes, Glymour, & Scheines 2001]. Section 3.4 describes more general tests of conditional independence. Such tests require specifying a significance level for the test, which is a user-specified parameter of the algorithm. Because the PC algorithm performs a sequence of tests without adjustment,

the significance level does not represent any (easily calculable) statistical feature of the output, but should only be understood as a parameter used to guide the search.

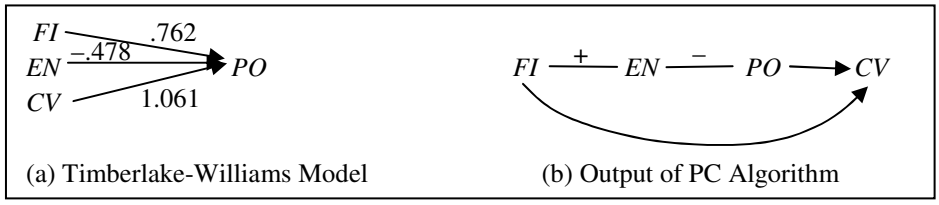
Assuming that the causal relations can be represented by a directed acyclic graph, the Causal Markov Assumption, the Causal Faithfulness Assumption, and consistent tests of conditional independence, in the large sample (i.i.d.) limit for a causally sufficient set of variables, the PC algorithm outputs a pattern that represents the true causal graph.

The PC algorithm has been shown to apply to very high dimensional data sets (under a stronger version of the Causal Faithfulness Assumption), both for finding causal structure [Kalisch & Buhlmann 2007] and for classification [Aliferis, Tsamardinos, & Statnikov 2003]. A version of the algorithm controlling the false discovery rate is available [Junning & Wang 2009].

3.1.1 Example - Foreign Investment

This example illustrates how the PC algorithm can find plausible alternatives to a model built from domain knowledge. Timberlake and Williams used regression to claim foreign investment in third-world countries promotes dictatorship [Timberlake & Williams 1984]. They measured political exclusion (*PO*) (i.e., dictatorship), foreign investment penetration in 1973 (*FI*), energy development in 1975 (*EN*), and civil liberties (*CV*) for 72 countries. Civil liberties was measured on an ordered scale from 1 to 7, with lower values indicating greater civil liberties.

Their inference is unwarranted. Their model (with the relations between the regressors omitted) and the pattern obtained from the PC algorithm using a .12 significance level to test for vanishing partial correlations) are shown in Figure 3.[1] We typically run the algorithms at a variety of different significance levels, and compare the results to see if any of the features of the output are constant.



(a) Timberlake-Williams Model (b) Output of PC Algorithm

Figure 3: Two Models of Foreign Investment

The PC Algorithm will not orient the *FI* – *EN* and *EN* – *PO* edges, and assumes that the edges are not due to an unmeasured common cause. Maximum likelihood estimates of any linear, Gaussian parameterization of any DAG represented by the pattern output by the PC algorithm requires that the influence of *FI* on *PO* (if any) be negative, and the models easily pass a likelihood ratio test. If any of these SEMs is correct, Timberlake and William's regression model appears to be a case in which an effect of the outcome variable is taken as a regressor.

Given the small sample size, and the uncertainty about the distributional assumptions, we do not present the alternative models suggested by the PC algorithm as

[1]Searches at lower significance levels remove the adjacency between FI and EN.

particularly well supported by the evidence. However, we do think that they are at least as well supported as the regression model, and hence serve to cast doubt upon conclusions drawn from that model.

3.1.2 Example - Spartina Biomass

This example illustrates a case where the PC algorithm output received some experimental confirmation. A recent textbook on regression [Rawlings 1988] skillfully illustrates regression principles and techniques for a biological study from a dissertation [Linthurst 1979] in which it is reasonable to think there is a causal process at work relating the variables. The question at issue is plainly causal: among a set of 14 variables, which have the most influence on an outcome variable, the biomass of Spartina grass? Since the example is the principle application given for an entire textbook on regression, the reader who reaches the 13th chapter may be surprised to find that the methods yield almost no useful information about that question.

According to Rawlings, Linthurst obtained five samples of Spartina grass and soil from each of nine sites on the Cape Fear Estuary of North Carolina. Besides the mass of Spartina (*BIO*), fourteen variables were measured for each sample:

- Free Sulfide (H_2S)
- Salinity (*SAL*)
- Redox potentials at pH 7 (EH_7)
- Soil pH in water (*PH*)
- Buffer acidity at pH 6.6 (*BUF*)
- Phosphorus concentration (*P*)
- Potassium concentration (*K*)
- Calcium concentration (*CA*)
- Magnesium concentration (*MG*)
- Sodium concentration (*NA*)
- Manganese concentration (*MN*)
- Zinc concentration (*ZN*)
- Copper concentration (*CU*)
- Ammonium concentration (NH_4)

The aim of the data analysis was to determine for a later experimental study which of these variables most influenced the biomass of Spartina in the wild. Greenhouse experiments would then try to estimate causal dependencies out in the wild. In the best case one might hope that the statistical analyses of the observational study would correctly select variables that influence the growth of Spartina in the greenhouse. In the worst case, one supposes, the observational study would find the wrong causal structure, or would find variables that influence growth in the wild (e.g., by inhibiting or promoting growth of a competing species) but have no influence in the greenhouse.

Using the SAS statistical package, Rawlings analyzed the variable set with a multiple regression and then with two stepwise regression procedures from the SAS package. A search through all possible subsets of regressors was not carried out, presumably because the candidate set of regressors is too large. The results were as follows:

(i) a multiple regression of *BIO* on all other variables gives only *K* and *CU* significant regression coefficients;

(ii) two stepwise regression procedures[2] both yield a model with *PH*, *MG*, *CA* and *CU* as the only regressors, and multiple regression on these variables alone gives them all significant coefficients;

(iii) simple regressions one variable at a time give significant coefficients to *PH*, *BUF*, *CA*, *ZN* and NH_4.

What is one to think? Rawling's reports that "None of the results was satisfying to the biologist; the inconsistencies of the results were confusing and variables expected to be biologically important were not showing significant effects." (p. 361).

This analysis is supplemented by a ridge regression, which increases the stability of the estimates of coefficients, but the results for the point at issue--identifying the important variables--are much the same as with least squares. Rawlings also provides a principal components factor analysis and various geometrical plots of the components. These calculations provide no information about which of the measured variables influence Spartina growth.

Noting that *PH*, for example, is highly correlated with *BUF*, and using *BUF* instead of *PH* along with *MG*, *CA* and *CU* would also result in significant coefficients, Rawlings effectively gives up on this use of the procedures his book is about:

> Ordinary least squares regression tends either to indicate that none of the variables in a correlated complex is important when all variables are in the model, or to arbitrarily choose one of the variables to represent the complex when an automated variable selection technique is used. A truly important variable may appear unimportant because its contribution is being usurped by variables with which it is correlated. Conversely, unimportant variables may appear important because of their associations with the real causal factors. It is particularly dangerous in the presence of collinearity to use the regression results to impart a "relative importance," whether in a causal sense or not, to the independent variables. (p. 362)

Rawling's conclusion is correct in spirit, but misleading and even wrong in detail. If we apply the PC algorithm to the Linthurst data then there is one robust conclusion*:* the only variable that may *directly* influence biomass in this population[3] is *PH*; *PH* is distinguished from all other variables by the fact that the correlation of every other variable (except *MG*) with *BIO* vanishes or vanishes when *PH* is conditioned on.[4] The relation is not symmetric; the correlation of *PH* and *BIO*, for example, does not vanish when *BUF* is controlled. The algorithm finds *PH* to be the only variable adjacent to *BIO*

---

[2]The "maximum R-square" and "stepwise" options in PROC REG in the SAS program.

[3]Although the definition of the population in this case is unclear, and must in any case be drawn quite narrowly.

[4]More exactly, at .05, with the exception of MG the partial correlation of every regressor with BIO vanishes when some set containing *PH* is controlled for; the correlation of *MG* with *BIO* vanishes when *CA* is controlled for.

no matter whether we use a significance level of .05 to test for vanishing partial correlations, or a level of 0.1, or a level of 0.2. In all of these cases, the PC algorithm (and the FCI algorithm, which allows for the possibility of latent variables in section 4.2) yields the result that *PH* and only *PH* can be directly connected with *BIO*. If the system is linear normal and the Causal Markov Assumption obtains, then in this population any influence of the other regressors on *BIO* would be blocked if *PH* were held constant. Of course, over a larger range of values of the variables there is little reason to think that *BIO* depends linearly on the regressors, or that factors that have no influence in producing variation within this sample would continue to have no influence.

Although the analysis cannot conclusively rule out possibility that *PH* and *BIO* are confounded by one or more unmeasured common causes, in this case the principles of the theory and the data argue against it. If *PH* and *BIO* have a common unmeasured cause *T*, say, and any other variable, Z_i, among the 13 others either causes *PH* or has a common unmeasured cause with *PH* (

Figure 4, in which we do not show connections among the **Z** variables), then Z_i and *BIO* should be correlated conditional on *PH*, which is statistically not the case.

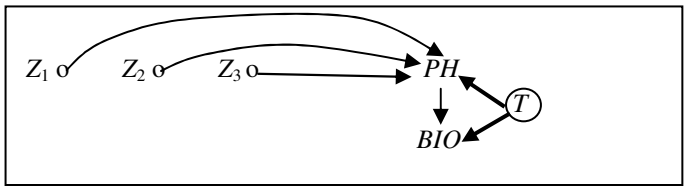


Figure 4 : *PH* and *BIO* Confounding?

The program and theory lead us to expect that if *PH* is forced to have values like those in the sample--which are almost all either below *PH* 5 or above *PH* 7-- then manipulations of other variables within the ranges evidenced in the sample will have no effect on the growth of Spartina. The inference is a little risky, since growing plants in a greenhouse under controlled conditions may not be a direct manipulation of the variables relevant to growth in the wild. If, for example, in the wild variations in *PH* affect Spartina growth chiefly through their influence on the growth of competing species not present in the greenhouse, a greenhouse experiment will not be a direct manipulation of *PH* for the system.

The fourth chapter of Linthurst's thesis partly confirms the PC algorithm's analysis. In the experiment Linthurst describes, samples of Spartina were collected from a salt marsh creek bank (presumably at a different site than those used in the observational study). Using a 3 x 4 x 2 (*PH* x *SAL* x *AERATION*) randomized complete block design with four blocks, after transplantation to a greenhouse the plants were given a common nutrient solution with varying values *PH* and *SAL* and *AERATION*. The *AERATION* variable turned out not to matter in this experiment. Acidity values were *PH* 4, 6 and 8. *SAL* for the nutrient solutions was adjusted to 15, 25, 35 and 45 ‰.

Linthurst found that growth varied with *SAL* at *PH* 6 but not at the other *PH* values, 4 and 8, while growth varied with *PH* at all values of *SAL* (p. 104). Each variable was

correlated with plant mineral levels. Linthurst considered a variety of mechanisms by which extreme *PH* values might control plant growth:

> At pH 4 and 8, salinity had little effect on the performance of the species. The pH appeared to be more dominant in determining the growth response. However, there appears to be no evidence for any causal effects of high or low tissue concentrations on plant performance unless the effects of pH and salinity are also accounted for. (p.108)
> The overall effect of pH at the two extremes is suggestive of damage to the root, thereby modifying its membrane permeability and subsequently its capacity for selective uptake. (p. 109).

A comparison of the observational and experimental data suggests that the PC Algorithm result was essentially correct and can be extrapolated through the variation in the populations sampled in the two procedures, but cannot be extrapolated through *PH* values that approach neutrality. The result of the PC search was that in the non-experimental sample, observed variations in aerial biomass were perhaps caused by variations in *PH*, but were not caused (at least not directly, relative to *PH*) by variations in other variables. In the observational data Rawlings reports (p. 358) almost all *SAL* measurements are around 30--the extremes are 24 and 38. Compared to the experimental study rather restricted variation was observed in the wild sample. The observed values of *PH* in the wild, however, are clustered at the two extremes; only four observations are within half a *PH* unit of 6, and no observations at all occurred at *PH* values between 5.6 and 7.1. For the observed values of *PH* and *SAL*, the experimental results appear to be in very good agreement with our results from the observational study: small variations in *SAL* have no effect on Spartina growth if the *PH* value is extreme.

3.1.3 College Plans

Sewell and Shah [Sewell & Shah 1968] studied five variables from a sample of 10,318 Wisconsin high school seniors.[5] The variables and their values are:

- *SEX* male = 0, female = 1
- *IQ* = Intelligence Quotient, lowest = 0, highest = 3
- *CP* = college plans yes = 0, no = 1
- *PE* = parental encouragement low = 0, high = 1
- *SES* = socioeconomic status lowest = 0, highest = 3

The question of interest is what the causes of college plans are. This data set is of interest because it has been used by a variety of different search algorithms that make different assumption. The different results illustrate the role that the different assumptions make in the output and are discussed in subsequent sections.

[5]Examples of the analysis of the Sewell and Shah data using Bayesian networks are given in Spirtes et al. (2001), and Heckerman (1998).

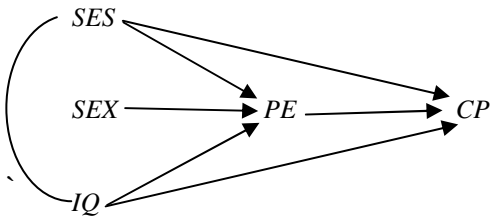


Figure 5: Model of Causes of College Plans

The pattern produced as the output of the PC algorithm is shown in Figure 5. The model predicts that *SEX* affects *CP* only indirectly via *PE*.

It is possible to predict the effects of some manipulations from the pattern, but not others. For example, because the pattern is compatible both with $SES \rightarrow IQ$ and with $SES \leftarrow IQ$, it is not possible to determine if *SES* is a cause or an effect of *IQ*, and hence it is not possible to predict the effect of manipulating *SES* on *IQ* from the pattern. On the other hand, it can be shown that all of the models in the conditional independence equivalence class represented by the pattern entail the same predictions about the quantitative effects of manipulating *PE* on *CP*. When *PE* is manipulated, in the manipulated distribution: $P(CP=0|PE=0) = .095$; $P(CP=1|PE=0) = .905$; $P(CP=0|PE=1) = .484$; $P(CP=1PE=1) = .516$ [Spirtes, Scheines, Glymour, & Meek 2004].

3.2 Greedy Equivalence Search Algorithm

Algorithms that maximize a score have certain advantages over constraint-based algorithms such as PC. When the data are not Gaussian, but the system is linear, they can take advantage of non-Gaussian distribution features. Extensive unpublished simulations find that at least one such algorithm, the Greedy Equivalence Search (GES) algorithm [Meek 1997] outperforms PC in such circumstances. GES can be used with a number of different scores for patterns, including posterior probabilities (for some parametric families and under some priors), and the Bayesian Information Criterion (BIC), which is an approximation of a class of posterior distributions in the large sample limit. The BIC score [Schwarz 1978] is: $-2 \ln(ML) + k \ln(n)$, where ML is the likelihood of the data at the maximum likelihood estimate of the parameters, k is the dimension of the model and n is the sample size. For uniform priors on models and smooth priors on the parameters, the posterior probability conditional on the data is a monotonic function of BIC in the large sample limit. In the forward stage of the search, starting with an initial (possibly empty) pattern, at each stage GES selects the pattern that is the one-edge addition compatible with the current pattern and has the highest score. The forward stage continues until no further additions improve the score. Then a reverse procedure is followed that removes edges according to the same criterion, until no improvement is found. The computational and convergence advantages of the algorithm depend on the fact that it searches over Markov equivalence classes of DAGs rather than individual

DAGs, and that only one forward stage and one backward stage are required for an asymptotically correct search. In the large sample limit, GES identifies the Markov equivalence class of the true graph if the assumptions above are met [Chickering 2002].

GES has proved especially valuable in searches for latent structure (GESMIMBuild) and in searches with multiple data sets (IMaGES). Examples are discussed in sections 4.4 and 5.3 .

3.3 LiNGAM

Standard implementations of the constraint-based and score-based algorithms above usually assume that continuous variables have multivariate Gaussian distributions. This assumption is inappropriate in many contexts such as EEG analysis where variables are known to deviate from Gaussianity.

The LiNGAM (Linear Non-Gaussian Acyclic Model) algorithm [Shimizu, Hoyer, Hyvärinen, & Kerminen 2006] is appropriate specifically for cases where each variable in a set of measured variables can be written as a linear function of other measured variables plus an independent noise component, where at most one of the measured variables' noise components may be Gaussian. For example, consider the system with the causal graph shown in

Figure 6 and assume X, Y, and Z are determined as follows, where a, b, and c are real-valued coefficients and ε_x, ε_y, and ε_z are independent noise components of which at least two are non-Gaussian.

(1) $X = \varepsilon_x$
(2) $Y = aX + \varepsilon_y$
(3) $Z = bX + cY + \varepsilon_z$

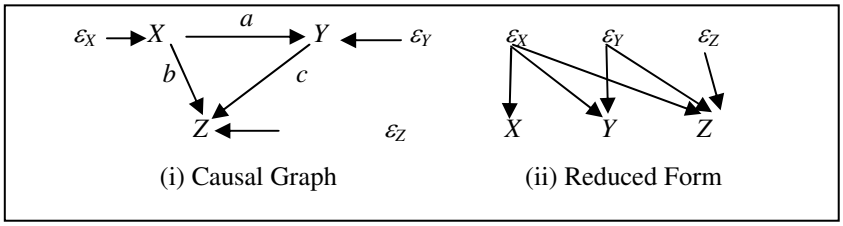


Figure 6: Causal Graph and Reduced Form

The equations can be rewritten in what economists called reduced form, also shown in Figure 6:

(4) $X = \varepsilon_X$
(5) $Y = a\varepsilon_X + \varepsilon_Y$
(6) $Z = b\varepsilon_X + ac\varepsilon_X + c\varepsilon_Y + \varepsilon_Z$

The standard Independent Components Analysis (ICA) procedure [Hyvärinen & Oja, 2000] can be used to recover a matrix containing the real-valued coefficients a, b, and c

from an i.i.d. sample of data generated from the above system of equations. The LiNGAM algorithm finds the correct matching of coefficients in this ICA matrix to variables and prunes away any insignificant coefficients using statistical criteria.

The procedure yields correct values even if the coefficients perfectly cancel, and the variables such as *X, Z* above are uncorrelated. Since coefficients are determined for each variable, we can always reconstruct the true unique DAG, instead of its Markov equivalence class. The procedure converges (at least) pointwise to the true DAG and coefficients assuming: (1) there are no unmeasured common causes; (2) the dependencies among measured variables are linear; (3) none of the relations among measured variables are deterministic; (4) i.i.d. sampling; (5) the Markov Condition; (6) at most one error or disturbance term is Gaussian. We do not know its complexity properties.

The LiNGAM procedure can be generalized to estimate causal relations among observables when there are latent common causes [Hoyer, Shimizu, & Kerminen 2006], although the result is not in general a unique DAG, and LiNGAM has been combined [Shimizu, Hoyer, & Hyvarinen 2009] with Silva's clustering procedure (section 4.4) for locating latent variables to estimate a unique DAG among latent variables, and also with search for cyclic graphs [Lacerda, Spirtes, Ramsey, & Hoyer 2008], and combined with the PC and GES algorithms when more than one disturbance term is Gaussian [Hoyer et al. 2008].

3.4 The kPC Algorithm

The kPC algorithm [Tillman, Gretton, & Spirtes, 2009] relaxes distributional assumptions further, allowing not only non-Gaussian noise with continuous variables, but also nonlinear dependencies. In many cases, kPC will return a unique DAG (even when there is more than one DAG in the Markov equivalence class. However, unlike LiNGAM there is no requirement that a certain number of variables be nonlinear or non-Gaussian.

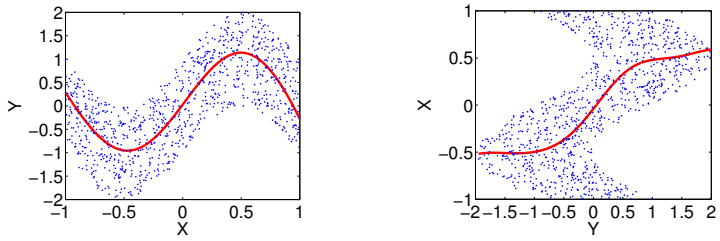
kPC consists of two stages. In the first stage of kPC, the standard PC algorithm is applied to the data using efficient implementations of the Hilbert-Schmidt Independence Criteria [Gretton, Fukumizu, Teo, Song, Scholkopf, & Smola, 2008], a nonparametric independence test and an extension of this test to the conditional cases based on the dependence measure given in [Fukumizu, Gretton, Sun, & Scholkopf, 2008]. This produces a pattern. Additional orientations are then possible if the true causal model, or a submodel (after removing some variables) of the true causal model is an *additive noise model* [Hoyer, Janzing, Mooij, Peters, & Scholkopf, 2009] that is *noninvertible*.

A set of variables is an additive noise model if (i) the function form of each variable can be expressed as a (possible nonlinear) smooth function of its parents in the true causal model plus an additive (Gaussian or non-Gaussian) noise component and (ii) the additive noise components are mutually independent. An additive noise model is noninvertible if we cannot reverse any edges in model and still obtain smooth functional forms for each variable and mutually independent additive noise components that fit the data.

For example, consider the two variable case where $X \rightarrow Y$ is the true DAG and we have the following function forms and additive noise components for X and Y:

$$X = \varepsilon_X,\ Y = \sin(\pi X) + \varepsilon_X,\ \ \varepsilon_X \sim Uniform(-1,1),\ \varepsilon_Y \sim Uniform(-1,1)$$

If we fit a nonparametric regression model for Y regressed on X, the forward model, Figure 7a, and for X regressed on Y, the backward model, Figure 7b, we observe $I(\hat{\varepsilon}_Y, X)$ and $\neg I(\hat{\varepsilon}_X, Y)$ since this additive noise model is noninvertible.



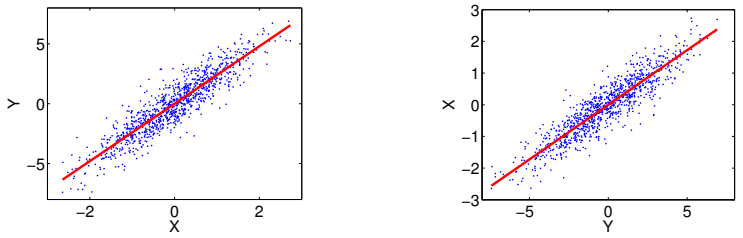
(a) **(b)**

Figure 7: Nonparametric regressions of (a) Y on X, and (b) X on Y with the data overlayed for nonlinear non-Gaussian case

Thus in this case, we can conclude that $X \to Y$ is the true DAG from the data since the additive noise model fits in only one direction, i.e. it is noninvertible. However, consider the following linear Gaussian case:

$$X = \varepsilon_X,\ Y = 2.4 \cdot X + \varepsilon_Y,\ \ \varepsilon_X \sim N(0,1),\ \varepsilon_Y \sim N(0,1)$$

After fitting nonparametric regression models for both directions, Figure 8, we find $I(\hat{\varepsilon}_Y, X)$ and $I(\hat{\varepsilon}_X, Y)$ so we cannot determine whether $X \to Y$ or $Y \to X$ is the correct DAG.



(a) **(b)**

Figure 8: Nonparametric regressions of (a) Y on X, and (b) X on Y with the data overlayed for linear Gaussian case

[Zhang and Hyvarinen, 2009] show that only a few special cases, other than the linear Gaussian case, exist where the additive noise model is invertible.

The second stage of kPC consists of searches for submodels that are consistent with the pattern learned in the first stage of kPC which may be noninvertible additive noise models. If such models are discovered, then further orientations of edges can be made resulting in an equivalence class of possible DAGs that is smaller than the Markov

equivalence class. In many cases, only a few variables need be nonlinear or non-Gaussian to obtain a unique DAG using kPC.

kPC requires the following additional assumption:

Weak Additivity Assumption: If the relationship between X and **Parents**(G,X) in the true DAG G cannot be expressed as a noninvertible additive noise model, there does not exist a Y in **Parents**(G,X) and alternative DAG G' such that Y and **Parents**(G',Y) can be expressed as a noninvertible additive noise model where X is included in **Parents**(G',Y).

This assumption does rule out invertible additive noise models or many cases where noise may not be additive, only the hypothetical case where we can fit an additive noise model to the data, but only in the incorrect direction. Weak additivity can be considered an extension of the simplicity intuitions underlying the causal faithfulness assumption, i.e. a complicated true model will not generate data resembling a different simpler model. Faithfulness can fail, but under a broad range of distributions, violations are Lebesgue measure zero [Spirtes, Glymour, & Scheines 2000]. Whether a similar justification can be given for the weak additivity assumption is an open question.

kPC is both correct and complete, i.e. it converges to the correct DAG or smallest possible equivalence class of DAGs in the limit under weak additivity and the assumptions of the PC algorithm.

3.4.1 Example - Auto MPG

Figure 9 shows the structures learned for the Auto MPG dataset, which records *MPG* fuel consumption of 398 automobiles in 1983 with 8 characteristics from the UCI database (Asuncion & Newman, 2007). The nominal variables *Year* and *Origin* were excluded.

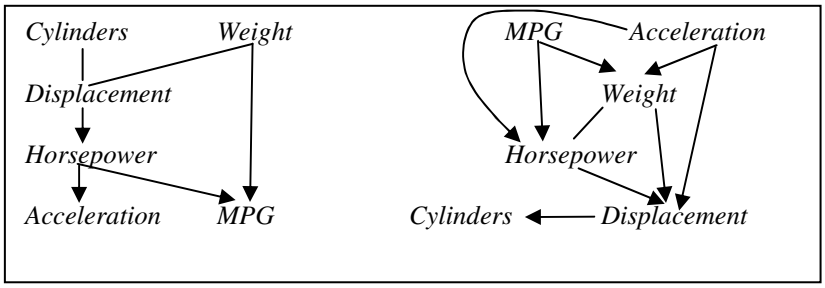


Figure 9: Automobile Models

The PC result indicates *MPG* causes *Weight* and *Horsepower*, and *Acceleration* causes *Weight*, *Horsepower*, and *Displacement*, which are clearly false. kPC finds the more plausible chain *Displacement* → *Horsepower* → *Acceleration* and finds *Horsepower* and *Weight* cause *MPG*.

3.4.2 Example - Forest Fires

The Forest Fires dataset contains 517 recordings of meteorological for forest fires observed in northeast Portugal and the total area burned (*Area*) [Asuncion & Newman 2007]. We again exclude nominal variables *Month* and *Year*.

Figure **10** shows the structures learned by PC and kPC for this dataset. kPC finds every variable other than *Area* is a cause of *Area*, which is sensible since each of these variables were included in the dataset by domain experts as predictors which influence the total area burned by forest fires.

The PC structure, however, indicates that Area is not associated with any of the variables, which are all assumed to be predictors by experts.

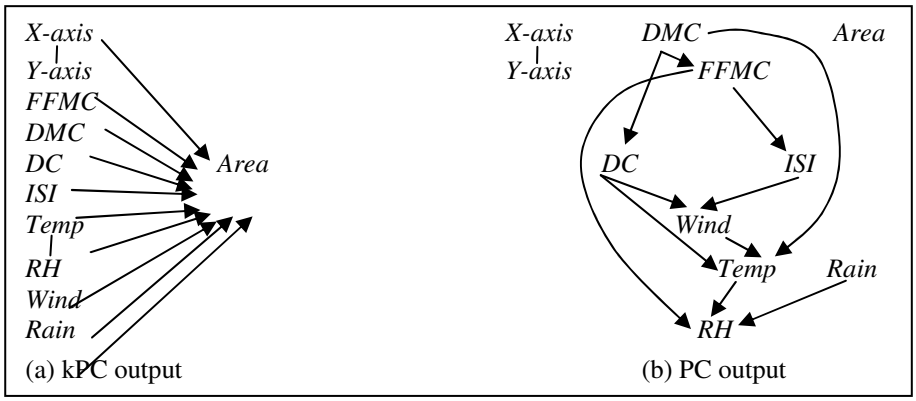


Figure 10: kPC and PC Forest Fires

4 Search For Latent Variable Models

The assumption that the observed variables are causally sufficient is usually unwarranted. In this section, we describe searches that do not make this assumption.

4.1 Distribution and Conditional Independence Equivalence

Let **O** be the set of observed variables, which may not be causally sufficient. If G_1 is a DAG over $\mathbf{V_1}$, G_2 is a DAG over $\mathbf{V_2}$, $\mathbf{O} \subseteq \mathbf{V_1}$, and $\mathbf{O} \subseteq \mathbf{V_2}$, G_1 and G_2 are **O**-conditional independence equivalent, if they both entail the same set of conditional independence relations among the variables in **O** (i.e. they have the same d-separation relations among the variables in **O**). $<G_1,\Theta_1>$ and $<G_2,\Theta_2>$ are **O-distribution equivalent** with respect to the parametric families Θ_1 and Θ_2 if and only if they represent the same set of marginal distributions over **O**.

It is possible that two directed graphs are conditional independence equivalent, or even distributionally equivalent (relative to given parametric families) but are not **O**-distributionally equivalent (relative to the same parametric families), as long as at least one of them contains a latent variable. Although there are algebraic techniques that determine when two Bayesian networks with latent variables are **O**-distributionally equivalent for some parametric families, or find features common to an **O**-distributional equivalence class, known algorithms to do so are not computationally feasible [Geiger & Meek 1999] for models with more than a few variables. In addition, if an unlimited number of latent variables are allowed, the number of DAGs that are **O**-distributionally equivalent may be infinite. Hence, instead of searching for **O**-distribution equivalence

classes of models, we will describe how to search for **O**-conditional independence classes of models. This is not as informative as the computationally infeasible strategy of searching for **O**-distribution equivalence classes, but is nevertheless correct.

It is often far from intuitive what constitutes a complete set of graphs **O**-conditional independence equivalent to a given graph although algorithms for deciding this now exist [Ali, Richardson, & Spirtes 2009].

4.2 The Fast Causal Inference Algorithm

The PC algorithm gives an asymptotically correct representation of the conditional independence equivalence class of a DAG without latent variables by outputting a pattern that represents all of the features that the DAGs in the equivalence class have in common. The same basic strategy can be used without assuming causal sufficiency, but the rules for detecting adjacencies and orientations are much more complicated, so we will not describe them in detail. The FCI algorithm6 outputs an asymptotically correct representation of the O-conditional independence equivalence class of the true causal DAG (assuming the Causal Markov and Causal Faithfulness Principles), in the form of a graphical structure called a partial ancestral graph that represents some of the features that the DAGs in the equivalence class have in common. The FCI algorithm takes as input a sample, distributional assumptions, optional background knowledge (e.g. time order), and a significance level, and outputs a partial ancestral graph. Because the algorithm uses only tests of conditional independence among sets of observed variables, it avoids the computational problems involved in calculating posterior probabilities or scores for latent variable models.

Just as the pattern can be used to predict the effects of some manipulations, a partial ancestral graph can also be used to predict the effects of some distributions. Instead of calculating the effects of manipulations for which every member of the simple **O**-distribution equivalence class agree, we can calculate the effects only of those manipulations for which every member of the simple **O**-conditional independence equivalence agree. This will typically predict the effects of fewer manipulations than could be predicted given the simple **O**-distributional equivalence class (because a larger set of graphs have to make the same prediction), but the predictions made will still be correct.

Even though the set *S* of DAGs in a conditional independence over **O** equivalence class is infinite, it is still possible to extract the features that the members of *S* have in common. For example, every member of the conditional independence class over **O** that contains the DAG in Figure 11 has a directed path from *PE* to *CP* and no latent common cause of *PE* and *CP*. This is informative because even though the data do not help choose between members of the equivalence class, insofar as the data are evidence

[6]The FCI algorithm is similar to Pearl's IC* algorithm [Pearl 2000] in many respects, and uses concepts bases on IC*; however IC* is computationally and statistically feasible only for a few variables.

for the disjunction of the members in the equivalence class, they are evidence that *PE* is a cause of *CP*.

A partial ancestral graph is analogous to a pattern, and represents the features common to a conditional independence over **O** equivalence class. Figure 11 shows an example of a DAG and the corresponding partial ancestral graph over **O** = {*IQ*, *SES*, *PE*, *CP*, *SEX*}. Two variables *A* and *B* are adjacent in a PAG that represents a conditional independence over **O** equivalence class, when *A* and *B* are not entailed to be independent (i.e. they are d-connected) conditional on any subset of the variables in **O**\{*A*,*B*} for each DAG in the conditional independence over **O** equivalence class. The "–" endpoint of the $PE \rightarrow CP$ edge means that *PE* is an ancestor of *CP* in every DAG in the conditional independence over **O** equivalence class. The ">" endpoint of the $PE \rightarrow CP$ edges means that *CP* is not an ancestor of *PE* in any member of the conditional independence over **O** equivalence class. The "o" endpoint of the *SES* o–o *IQ* edge makes no claim about whether *SES* is an ancestor of *IQ* or not.

Applying the FCI algorithm to the Sewell and Shah data yields the PAG in Figure 11. The output predicts that when *PE* is manipulated, the following conditional probabilities hold: $P(CP=0|PE=0) = .063$; $P(CP=1|PE=0) = .937$; $P(CP=0|PE=1) = .572$; $P(CP=1PE=1) = .428$. These estimates are close to the estimates given by the output of the PC algorithm, although unlike the PC algorithm the output of the FCI algorithm posits the existence of latent variables. A bootstrap test of the output run at significance level 0.001 yielded the same results on 8 out of 10 samples. In the other two samples, the algorithm could not calculate the effect of the manipulation.

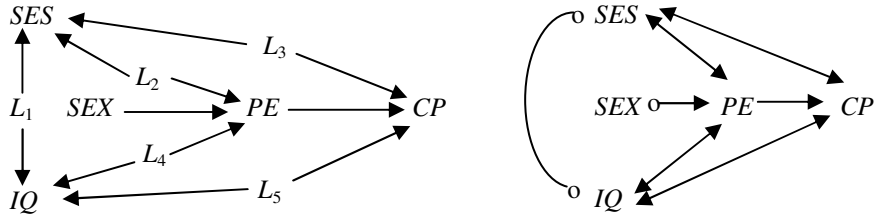


Figure 11: DAG and Partial Ancestral Graph

4.2.1 Online Course

This is an example where there was some experimental confirmation of the FCI causal model. Carnegie Mellon University offers a full semester online course that serves as a tutor on the subject of causal reasoning.[7] The course contains a number of different modules that contain both text and interactive online exercises that illustrate various concepts. Each module ends with a quiz that students must take. The interactive exercises are purely voluntary and play no role in calculating the student's final grade. It is possible to print the text from the online modules, but a student who studies from the printed text cannot use the online interactive exercises. The following variables were measured for each student:

[7]See http://oli.web.cmu.edu/openlearning/forstudents/freecourses/csr

- Pre-test (%)
- Print-outs (% modules printed)
- Quiz Scores (avg. %)
- Voluntary Exercises (% completed)
- Final Exam (%)
- 9 other variables

Using data from 2002, and some background knowledge about causal order, the output of the FCI algorithm was the PAG shown in

Figure **12**a. That model predicts that interventions that stops students from printing out the text and encourages students to use the online interactive exercises should raise the final grade in the class.

In 2003, students were advised that completing the voluntary exercises seemed to be important in helping grades, but that printing out the modules seemed to prevent completing the voluntary exercises. They were advised that, if they printed out the text they should make extra effort to go online and complete the interactive online exercises. Data on the same variables was gathered in 2003, and the output of the FCI algorithm is shown

Figure **12**b. The interventions to discourage printing and encourage the use of the online interactive exercises were largely successful, and the PAG output by the FCI algorithm from the 2003 data is exactly the PAG one would expect after intervening on the PAG output by the FCI algorithm from the 2002 data.

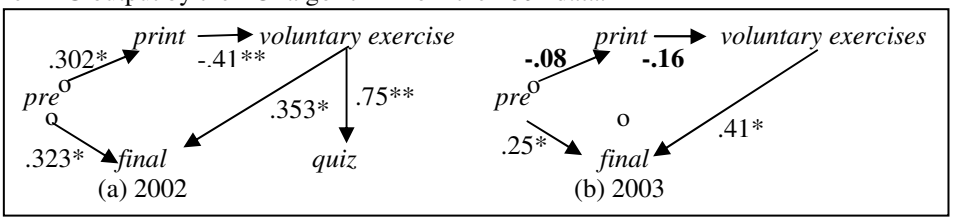


Figure 12: Online Course Printing

4.3 Errors in Variables: Combining Constraint Based Search and Bayesian Reasoning

In some cases the parameters of the output of the FCI algorithm are not identifiable or it is important to find not a particular latent variable model, but an equivalence class of latent variable models. In some of those cases the FCI algorithm can be combined with Bayesian methods.

4.3.1 Example - Lead and IQ

The next example shows how the FCI algorithm can be used to find a PAG, which can then be used as a starting point for a search for a latent variable DAG model and Bayesian estimation of parameters. It also illustrates how such a procedure produces different results than simply applying regression or using regression to generate more sophisticated models, such as errors-in-variables models.

By measuring the concentration of lead in a child's baby teeth, Herbert Needleman was the first epidemiologist to even approximate a reliable measure of cumulative lead exposure. His work helped convince the United States to eliminate lead from gasoline and most paint [Needleman 1979]. In their 1985 article in *Science* [Needleman, Geiger, & Frank 1985], Needleman, Geiger and Frank gave results for a multivariate linear regression of children's IQ on lead exposure. Having started their analysis with almost 40 covariates, they were faced with a variable selection problem to which they applied backwards-stepwise variable selection, arriving at a final regression model involving lead and five of the original 40 covariates. The covariates were measures of genetic contributions to the child's IQ (the parent's IQ), the amount of environmental stimulation in the child's early environment (the mother's education), physical factors that might compromise the child's cognitive endowment (the number of previous live births), and the parent's age at the birth of the child, which might be a proxy for many factors. The measured variables they used are as follows:

ciq - child's verbal IQ score
piq - parent's IQ scores
lead - measured concentration in baby teeth
mab - mother's age at child's birth
med - mother's level of education in years
fab - father's age at child's birth
nlb - number of live births previous to the sampled child

The standardized regression solution[8] is as follows, with t-ratios in parentheses. Except for *fab*, which is significant at 0.1, all coefficients are significant at 0.05, and $R^2 = .271$.

$$\hat{ciq} = -.143\ lead + .219\ med + .247\ piq + .237\ mab - .204\ fab - .159\ nlb$$
$$(2.32) \quad (3.08) \quad (3.87) \quad (1.97) \quad (1.79) \quad (2.30)$$

This analysis prompted criticism from Steve Klepper and Mark Kamlet, economists at Carnegie Mellon [Klepper, 1988/Klepper, Kamlet, & Frank 1993]. Klepper and Kamlet correctly argued that Needleman's statistical model (a linear regression) neglected to account for measurement error in the regressors. That is, Needleman's measured regressors were in fact imperfect proxies for the actual but latent causes of variations in IQ, and in these circumstances a regression analysis gives a biased estimate of the desired causal coefficients and their standard errors. Klepper and Kamlet constructed an errors-in-variables model to take into account the measurement error. See

Figure 13, where the latent variables are in boxes, and the relations between the regressors are unconstrained.

Unfortunately, an errors-in-variables model that explicitly accounts for Needleman's measurement error is "underidentified," and thus cannot be estimated by classical techniques without making additional assumptions. Klepper, however, worked out an ingenious technique to bound the estimates, provided one could reasonably bound the

[8] The covariance data for this reanalysis was originally obtained from Needleman by Steve Klepper, who generously forwarded it. In this, and all subsequent analyses described, the correlation matrix was used.

amount of measurement error contaminating certain measured regressors [Klepper, 1988; Klepper et al. 1993]. The required measurement error bounds vary with each problem, however, and those required in order to bound the effect of actual lead exposure below 0 in Needleman's model seemed wholly unreasonable. Klepper concluded that the statistical evidence for Needleman's hypothesis was indeed weak. A Bayesian analysis, based on Gibbs sampling techniques, found that several posteriors corresponding to different priors lead to similar results. Although the size of the Bayesian point estimate for lead's influence on IQ moved up and down slightly, its sign and significance (the 95% central region in the posterior over the *lead-iq* connection always included zero) were robust.

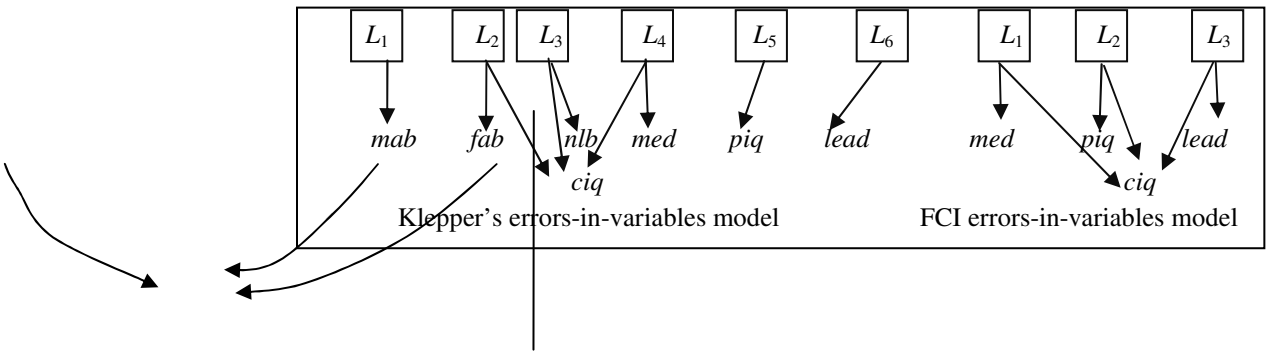


Figure 13: Errors-in-Variables Models

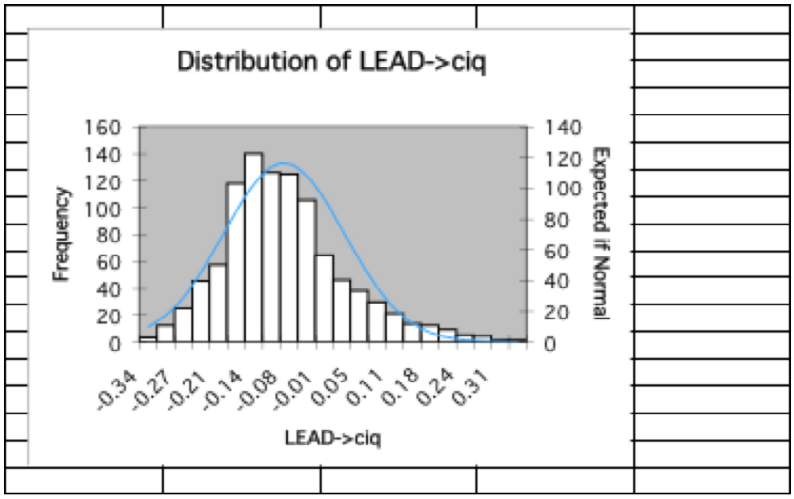


Figure 14: Posterior for Klepper's Model

A reanalysis using the FCI algorithm produced different results [Scheines 2000]. Scheines first used the FCI algorithm to generate a PAG, which was subsequently used as

the basis for constructing an errors-in-variables model. The FCI algorithm produced a PAG that indicated that *mab*, *fab*, and *nlb* are *not* adjacent to *ciq*, contrary to Needleman's regression.[9] If we construct an errors-in-variables model compatible with the PAG produced by the FCI algorithm, the model does not contain *mab*, *fab*, or *nlb*. See

Figure 13. (We emphasize that there are other models compatible with the PAG, which are not errors-in-variables models; the selection of an error-in-variables model from the set of models represented by the PAG is an assumption.) In fact the variables that the FCI algorithm eliminated were precisely those, which required unreasonable measurement error assumptions in Klepper's analysis. With the remaining regressors, Scheines specified an errors-in-variables model to parameterize the effect of actual lead exposure on children's' IQ. This model is still underidentified but under several priors, nearly all the mass in the posterior was over negative values for the effect of actual lead exposure (now a latent variable) on measured IQ. In addition, applying Klepper's bounds analysis to this model indicated that the effect of actual lead exposure on *ciq* was bounded below zero given reasonable assumptions about the degree of measurement error.

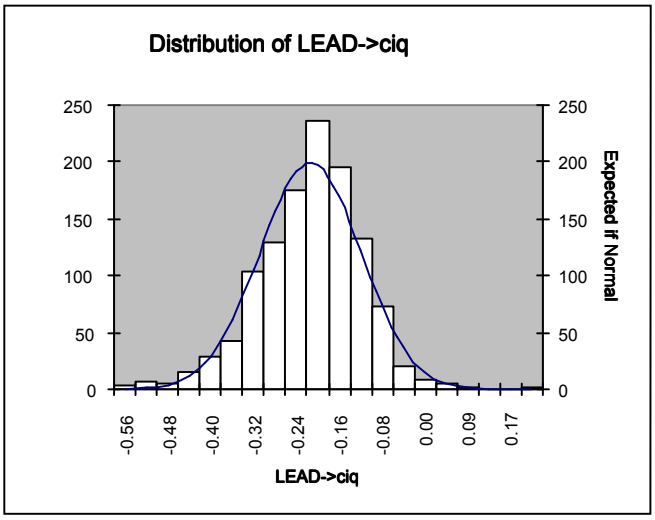


Figure 15: Posterior for FCI model

4.4 BuildPureClusters and MIMBuild

Searches using conditional independence constraints are correct, but completely uninformative for some common kinds of data sets. Consider the model *S* in Figure 16. The data comes from a survey of test anxiety indicators administered to 335 grade 12 male students in British Columbia [Gierl & Todd 1996]. The survey contains 20

[9] The fact that *mab* had a significant regression coefficient indicates that *mab* and *ciq* are correlated conditional on the other variables; the FCI algorithm concluded that *mab* is not a cause of *ciq* because *mab* and *ciq* are unconditionally uncorrelated.

measures of symptoms of anxiety under test conditions. Each question is about a symptom of anxiety. For example, question 8 is about how often one feels "jittery when taking tests". The answer is observed on a four-point approximately Likert scale (almost never, sometimes, often, or almost always). As in many such analyses, we will assume that the variables are approximately Gaussian.

Each *X* variable represents an answer to a question on the survey. For reasons to be explained later, not all of the questions on the test have been included in the model. There are three unobserved common causes in the model: *Emotionality*, *Care about achieving* (which will henceforth be referred to as *Care*) and *Self-defeating*. The test questions are of little interest in themselves; of more interest is what information they reveal about some unobserved psychological traits. If *S* is correct, there are no conditional independence relations among the *X* variables alone - the only entailed conditional independencies require conditioning on an unobserved common cause. Hence the FCI algorithm would return a completely unoriented PAG in which every pair of variables in **X** is adjacent. Such a PAG makes no predictions at all about the effects of manipulations of the observed variables.

Furthermore, in this case, the effects of manipulating the observed variables (answers to test questions) are of no interest - the interesting questions are about the effects of manipulating the unobserved variables and the qualitative causal relationships between them.

Although PAGs can reveal the existence of latent common causes (as by the double-headed arrows in Figure 11 for example), before one could make a prediction about the effect of manipulating an unobserved variable(s), one would have to identify what the variable (or variables) is, which is never possible from a PAG.

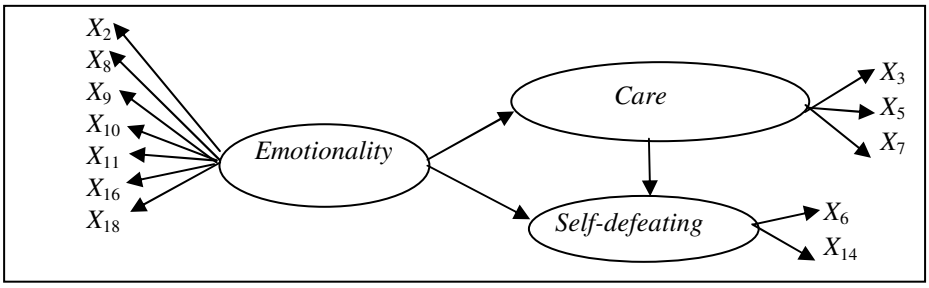


Figure 16: SEM S

Models such as *S* are *multiple indicator models*, and can be divided into two parts: the measurement model, which contains the edges between the unobserved variables and the observed variables (e.g. *Emotionality* $\rightarrow X_2$), and the structural model, which contains the edges between the unobserved variables (e.g. *Emotionality* $\rightarrow$ *Care*).

The **X** variables in *S* ($\{X_2, X_3, X_5, X_7, X_8, X_9, X_{10}, X_{11}, X_{14}, X_{16}, X_{18}\}$) were chosen with the idea that they indirectly measure some psychological trait that cannot be directly observed. Ideally, the **X** variables can be broken into clusters, where each variable in the cluster is caused by one unobserved cause common to the members of the cluster, and a

unique error term uncorrelated with the other error terms, and nothing else. From the values of the variables in the cluster, it is then possible to make inferences about the value of the unobserved common cause. Such a measurement model is called *pure*. In psychometrics, pure measurement models satisfy the property of local independence: each measured variable is independent of all other variables, conditional on the unobserved variable it measures. In Figure 16, the measurement model of *S* is pure.

If the measurement model is impure (i.e. there are multiple common causes of a pair of variables in **X**, or some of the **X** variables cause each other) then drawing inferences about the values of the common causes is much more difficult. Consider the set $\mathbf{X'} = \mathbf{X} \cup \{X_{15}\}$. If X_{15} indirectly measured (was a direct effect of) the unobserved variable *Care*, but X_{10} directly caused X_{15}, then the measurement model over the expanded set of variables would not be pure. If a measurement model for a set **X'** of variables is not pure, it is nevertheless possible that some subset of **X'**, such as **X**, has a pure measurement model. If the only reason that the measurement model is impure is that X_{10} causes X_{15} then $\mathbf{X} = \mathbf{X'} \backslash \{X_{15}\}$ does have a pure measurement model, because all the "impurities" have been removed. *S* does not contain all of the questions on the survey precisely because various tests described below indicated that they some of them needed to be excluded in order to have a pure measurement model.

The task of searching for a multiple indicator model can then be broken into two parts: first finding clusters of variables so that the measurement model is pure; second, use the pure measurement model to make inferences about the structural model.

Factor analysis is often used to determine the number of unmeasured common causes in a multiple indicator model, but there are important theoretical and practical problems in using factor analysis in this way. Factor analysis constructs models with unobserved common causes (factors) of the observed **X** variables. However, factor analysis models typically connect each unobserved common cause (factor) to each **X** variable, so the measurement model is not pure. A major difficulty with giving a causal interpretation to factor analytic models is that the observed distribution does not determine the covariance matrix among the unobserved factors. Hence, a number of different factor analytic models are compatible with the same observed data [Harman 1976]. In order to reduce the underdetermination of the factor analysis model by the data, it is often assumed that the unobserved factors are independent of each other; however, this is clearly not an appropriate assumption for unobserved factors that are supposed to represent actual causes that may causally interact with each other. In addition, simulation studies indicate that factor analysis is not a reliable tool for estimating the correct number of unobserved common causes [Glymour 1998].

On this data set, factor analysis indicates that there are 2 unobserved direct common causes, rather than 3 unobserved direct common causes [Bartholomew, Steele, Moustaki, & Galbraith 2002]. If a pure measurement model is constructed from the factor analytic model by associating each observed *X* variable only with the factor that it is most strongly associated with, the resulting model fails a statistical test (has a p-value of zero) [Silva, Scheines, Glymour, & Spirtes 2006]. A search for pure measurement models that depends upon testing vanishing tetrad constraints is an alternative to factor analysis.

Conceptually, the task of building a pure measurement model from the observed variables can be broken into 3 separate tasks:

1. Select a subset of the observed variables that form a pure measurement model.
2. Determine the number of clusters (i.e. the number of unobserved common causes) that the observed variables measure.
3. Cluster the observed variables into the proper groups (so each group has exactly one unobserved direct common cause.)

It is possible to construct pure measurement models using vanishing tetrad constraints as a guide [Silva et al. 2006]. A *vanishing tetrad constraint* holds among X, Y, Z, W when $\text{cov}(X,Y) \cdot \text{cov}(Z,W) - \text{cov}(X,Z) \cdot \text{cov}(Y,W) = 0$. A pure measurement model entails that each X_i variables is independent of every other X_j variable conditional on its unobserved parent, e.g. S entails X_2 is independent of X_j conditional on *Emotionality*. These conditional independence relations cannot be directly tested, because *Emotionality* is not observed. However, together with the other conditional independence relations involving unobserved variables entailed by S, they imply vanishing tetrad constraints on the observed variables that reveal information about the measurement model that does not depend upon the structural model among the unobserved common causes. The basic idea extends back to Spearman's attempts to use vanishing tetrad constraints to show that there was a single unobserved factor of intelligence that explained a variety of observed competencies [Spearman 1904].

Because X_2 and X_8 have one unobserved direct common cause (*Emotionality*), and X_3 and X_5 have a different unobserved direct common cause (*Care*), S entails $\text{cov}_S(X_2, X_3) \cdot \text{cov}_S(X_5, X_8) = \text{cov}_S(X_2, X_5) \cdot \text{cov}_S(X_3, X_8) \neq \text{cov}_S(X_2, X_8) \cdot \text{cov}_S(X_3, X_5)$ for all values of the model's free parameters (here cov_S is the covariance matrix entailed by S).[10] On the other hand, because X_2, X_8, X_9, and X_{10} all have one unobserved common cause (*Emotionality*) as a direct common cause, the following vanishing tetrad constraints are entailed by S: $\text{cov}_S(X_2, X_8) \cdot \text{cov}_S(X_9, X_{10}) = \text{cov}_S(X_2, X_9) \cdot \text{cov}_S(X_8, X_{10}) = \text{cov}_S(X_2, X_{10}) \cdot \text{cov}_S(X_8, X_9)$ [Spirtes et al. 2001]. The BuildPureClusters algorithm uses the vanishing tetrad constraints as a guide to the construction of pure measurement models, and in the large sample limit reliably succeeds if there is a pure measurement model among a large enough subset of the observed variables [Silva et al. 2006].

In this example, BuildPureClusters automatically constructed the measurement model corresponding to the measurement model of S. The clustering on statistical grounds makes substantive sense, as indicated by the fact that it is similar to a theory-based clustering based on background knowledge about the content of the questions; however BuildPureClusters removes some questions, and splits one of the clusters of questions constructed from domain knowledge into two clusters.

---

[10] The inequality is based on an extension of the Causal Faithfulness Assumption that states that vanishing tetrad constraints that are not entailed for all values of the free parameters by the true causal graph are assumed not to hold.

Once a pure measurement model has been constructed, there are several algorithms for finding the structural model. One way is to estimate the covariances among the unobserved common causes, and then input the estimated covariances to the FCI algorithm. The output is then a PAG among the unobserved common causes. Alternative searches for the structural model include the MIMBuild and GESMIMBuild algorithms, which output patterns [Silva et al. 2006].

In this particular analysis, the MIMBuild algorithm, which also employs vanishing tetrad constraints, was used to construct a variety of output patterns corresponding to different values of the search parameters. The best pattern returned contains an undirected edge between every pair of unobserved common causes. (*S* is an example that is compatible with the pattern, but any other orientation of the edges among the three unobserved common causes that does not create a cycle is also compatible with the pattern.) The resulting model (or set of models) passes a statistical test with a p-value of 0.47.

4.4.1 Example - Religion and Depression

This example shows how an automated causal search produces a model that is compatible with background knowledge, but fits much better than a model that was built from theories about the domain.

Bongjae Lee from the University of Pittsburgh organized a study to investigate religious/spiritual coping and stress in graduate students [Silva & Scheines 2004]. In December of 2003, 127 Masters in Social Works students answered a questionnaire intended to measure three main factors:

- *Stress*, measured with 21 items, each using a 7-point scale (from "not all stressful" to "extremely stressful") according to situations such as: "fulfilling responsibilities both at home and at school"; "meeting with faculty"; "writing papers"; "paying monthly expenses"; "fear of failing"; "arranging childcare";
- *Depression*, measured with 20 items, each using a 4-point scale (from "rarely or none" to "most or all the time") according to indicators as: "my appetite was poor"; "I felt fearful"; "I enjoyed life" "I felt that people disliked me"; "my sleep was restless";
- *Spiritual coping*, measured with 20 items, each using a 4-point scale (from "not at all" to "a great deal") according to indicators such as: "I think about how my life is part of a larger spiritual force"; "I look to God (high power) for strength in crises"; "I wonder whether God (high power) really exists"; "I pray to get my mind off of my problems";

The goal of the original study was to use graphical models to quantify how *Spiritual coping* moderates the association of *Stress* and *Depression*, and hypothesized that *Spiritual coping* reduces the association of *Stress* and *Depression*. The theoretical model (Figure 17) fails a chi-square test: $p = 0$. The measurement model produced by BuildPureClusters is shown in Figure 18. Note that the variables selected automatically are proper subsets of Lee's substantive clustering. The full model automatically produced with GESMIMBuild with the prior knowledge that *Stress* is not an effect of other latent

variables is given in Figure 19. This model passes a chi square test, p = 0.28, even though the algorithm itself does not try to directly maximize the fit. Note that it supports the hypothesis that *Depression* causes *Spiritual Coping* rather than the other way around. Although this conclusion is not conclusive, the example does illustrate how the algorithm can find a theoretically plausible alternative model that fits the data well.

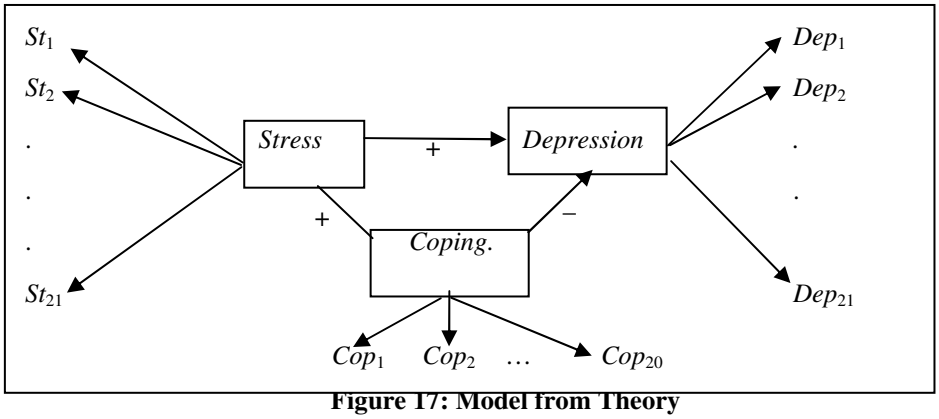


Figure 17: Model from Theory

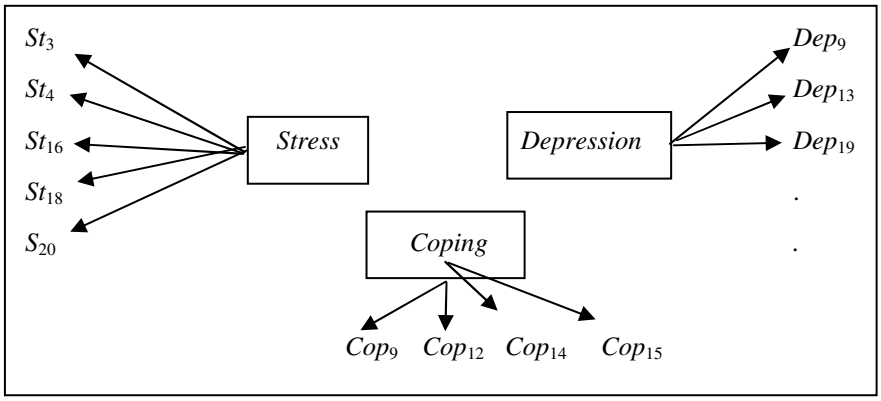


Figure 18: Output of BuildPureClusters

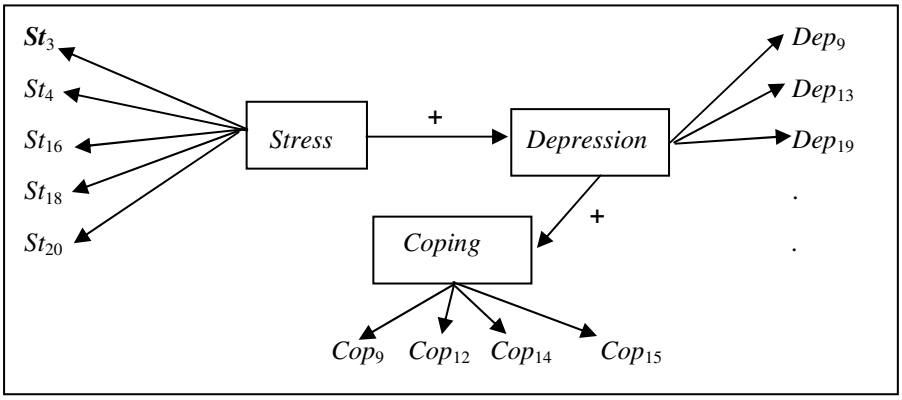


Figure 19: Output of GESMIMBuild

5 Time Series and Feedback

The models described so far are for "equilibrium." That is, they assume that an intervention fixes the values of a variable or variables, and that the causal process results in stable values of effect variables, so that time can be ignored. When time cannot be ignored, representation, interventions and search are all more complicated.

Time series models with a causal interpretation are naturally represented by directed acyclic graphs in at least three different forms: A graph whose variables are indexed by time, a "unit" graph giving a substructure that is repeated in the time indexed graph, and a finite graph that may be cyclic. Models of the first kind have been described as "Dynamical Causal Models" but the description does not address the difficulties of search. Pursuing a strategy of the PC or FCI kind, for example, requires a method of correctly estimating conditional independence relations.

5.1 Time series models

Chu and Glymour [2008] describe conditional independence tests for additive models, and use these tests in a slight modification of the PC and FCI algorithms. The series data is examined by standard methods to determine the requisite number of lags. The data are then replicated a number of times equal to the lags, delaying the first replicant by one time step, the second by two time steps, and so on, and conditional independence tests applied to the resulting sets of data. They illustrate the algorithm with climate data.

Climate teleconnections are associations of geospatially remote climate phenomena produced by atmospheric and oceanic processes. The most famous, and first established teleconnection, is the association of El Nino/Southern Oscillation (*ENSO*) with the failure of monsoons in India. A variety of associations have been documented among sea surface temperatures (*SST*), atmospheric pressure at sea level (*SLP*), land surface temperatures (*LST*) and precipitation over land areas. Since the 1970s data from a sequence of satellites have provided monthly (and now daily) measurements of such variables, at resolutions as small as 1 square kilometer. Measurements in particular spatial regions have been clustered into time-indexed indices for the regions, usually by principal components analysis, but also by other methods. Climate research has established that some of these phenomena are exogenous drivers of others, and has sought physical mechanisms for the teleconnections.

Chu and Glymour (2008) consider data from the following 6 ocean climate indices, recorded monthly from 1958 to 1999, each forming a time series of 504 time steps:

- *QBO* (*Quasi Biennial Oscillation*): Regular variation of zonal stratospheric winds above the equator
- *SOI* (*Southern Oscillation*): Sea Level Pressure (*SLP*) anomalies between Darwin and Tahiti
- *WP* (*Western Pacific*): Low frequency temporal function of the 'zonal dipole' *SLP* spatial pattern over the North Pacific.

- *PDO* (*Pacific Decadal Oscillation*): Leading principal component of monthly Sea Surface Temperature (*SST*) anomalies in the North Pacific Ocean, poleward of 20° N
- *AO* (*Arctic Oscillation*): First principal component of SLP poleward of 20° N
- *NAO* (*North Atlantic Oscillation*) Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland

Some connections among these variables are reasonably established, but are not assumed in the analysis that follows. In particular, *SO* and *NAO* are thought to be exogenous drivers.

After testing for stationarity, the PC algorithm yields the structure for the climate data shown in Figure 20. The double-headed arrows indicate the hypothesis of common unmeasured causes.[11] So far as the exogenous drivers are concerned, the algorithm output is in accord with expert opinion.

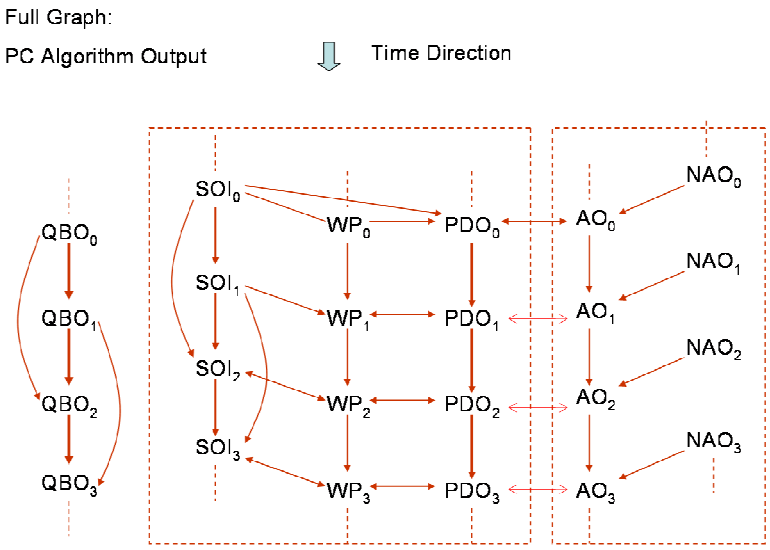


Figure 20: Climate Time Series

Monthly time series of temperatures and pressures at the sea surface present a case in which one might think that the causal processes take place more rapidly that the sampling rate. If so, then the causal structure in between time samples, the "contemporaneous" causal structure, should look much like a unit of the time series causal structure. When we sample at intervals of time as in economic, climate, and other time series, can we discover what goes on in the intervals between samples? Swanson and Granger suggested that an autoregression be used to remove the effects on each variable of

[11] When under the usual assumptions, the PC algorithm produces double headed arrows, they reliably indicate common unobserved causes as will FCI. But unlike FCI, PC is not complete in this respect.

variables at previous times, and a search could then be applied to the residual correlations [Swanson & Granger 1997]. The search they suggested was to assume a chain and to test it by methods described in [Glymour, Scheines, Spirtes, & Kelly 1987], some of the work whose aims and methods Cartwright previously sought to demonstrate is impossible [Cartwright 1994]. But a chain model of contemporaneous causes is far too special a case. Hoover, and later, Moneta & Spirtes, proposed applying PC to the residuals [Hoover & Demiralp 2003; Moneta & Spirtes 2006]. (Moneta also worked out the statistical corrections to the correlations required by the fact that they are obtained as residuals from regressions.) When that is done for model above, the result is the unit structure of the time series: $QBO \quad SOI \rightarrow WP \leftrightarrow PDO \leftrightarrow AO \leftarrow NA$

5.2 Cyclic Graphs

Since the 1950s, the engineering literature has developed methods for analyzing the statistical properties of linear systems described by cyclic graphs. The literature on search is more recent. Spirtes showed that linear systems with independent noises satisfy a simple generalization of d-separation, and the idea of faithfulness is well-defined for such systems [Spirtes 1995]; Pearl & Dechter extended these results to discrete variable systems [Pearl & Dechter 1996]. Richardson proved some of the essential properties of such graphs, and developed a pointwise consistent PC style algorithm for search [Richardson 1996]. More recently, an extension of the LiNGAM algorithm for linear, cyclic, non-Gaussian models has been developed [Lacerda et al. 2008].

5.3 Distributed Multiple Data Sets: ION and IMaGES

Data mining has focused on learning from a single database, but inferences from multiple databases are often needed in social science, multiple subject time series in physiological and psychological experiments, and to exploit archived data in many subjects. Such data sets typically pose missing variable problems: some of what is measured in one study or for one subject, may not be measured in another. In many cases such multiple data sets cannot, for physical, sociological or statistical reasons, be merged into a single data set with missing variables. There are two strategies for this kind of problem: learn a structure or set of structures separately for each data set and then find the set of structures consistent with the several "marginal" structures, or learn a single set of structures by evaluating steps in a search procedure using all of the data sets. The first strategy could be carried out using PC, kPC GES, FCI, LiNGAM or other procedure on each data set, and then using an algorithm that returns a description of the set of all graphs, or mixed graphs, consistent with the results from each database [Tillman, Danks, & Glymour 2008]. Tillman, Danks and Glymour have used such a procedure in combination with GES and FCI. The result in some (surprising) cases is a unique partial ancestral graph, and in other cases a large set of alternatives collectively carrying little information. The second strategy has been implemented in the IMaGES algorithm by Ramsey, et al. (submitted). The algorithm uses GES, but at each step in the evaluation of a candidate edge addition or removal, the candidate is scored separately by BIC on each data set and the average of the BIC scores is used by the algorithm in edge addition or

deletion choices. The IMaGES strategy is more limited—no consistency proof is available when the samples are from mixed distributions, and a proof of convergence of averages of BIC scores to a function of posteriors is only available when the sample sizes of several data sets are equal. Nonetheless, IMaGES has been applied to fMRI data from multiple subjects with remarkably good results. For example, an fMRI study of responses to visually presented rhyming and non-rhyming words and non-words should produce a left hemisphere cascade leading to right hemisphere effects, which is exactly what IMaGES finds, using only the prior knowledge that the input variable is not an effect of other variables.

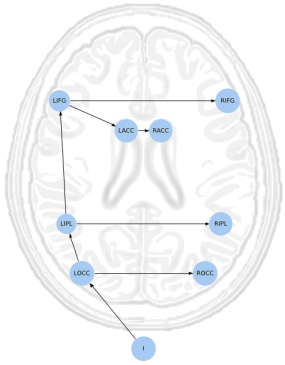


Figure 21: IMaGES Output for fMRI Data

6 Conclusion

The discovery of d-separation, and the development of several related notions, has made possible principled search for causal relations from observational and quasi-experimental data in a host of disciplines. New insights, algorithms and applications have appeared almost every year since 1990, and they continue. We are seeing a revolution in understanding of what is and is not possible to learn from data, but the insights and methods have seeped into statistics and applied science only slowly. We hope that pace will quicken.

7 Appendix

A *directed graph* (e.g. G_1 of

Figure 22) consists of a set of vertices and a set of directed edges, where each edge is an ordered pair of vertices. In G_1, the vertices are $\{A,B,C,D,E\}$, and the edges are $\{B \rightarrow A, B \rightarrow C, D \rightarrow C, C \rightarrow E\}$. In G_1, B is a *parent* of A, A is a *child* of B, and A and B are *adjacent* because there is an edge $A \rightarrow B$. A *path* in a directed graph is a sequence of adjacent edges (i.e. edges that share a single common endpoint). A *directed path* in a directed graph is a sequence of adjacent edges all pointing in the same direction. For example, in G_1, $B \rightarrow C \rightarrow E$ is a directed path from B to E. In contrast, $B \rightarrow C \leftarrow D$ is a path, but not a directed path in G_1 because the two edges do not point in the same direction; in addition, C is a *collider* on the path because both edges on the path are directed into C. A triple of vertices $<B,C,D>$ is a *collider* if there are edges $B \rightarrow C \leftarrow D$

in G_1; <*B*,*C*,*D*> is an *unshielded collider* if in addition there is no edge between *B* and *D*. *E* is a *descendant* of *B* (and *B* is an *ancestor* of *E*) because there is a directed path from *B* to *E*; in addition, by convention, each vertex is a descendant (and ancestor) of itself. A directed graph is *acyclic* when there is no directed path from any vertex to itself: in that case the graph is a directed acyclic graph, or DAG for short.

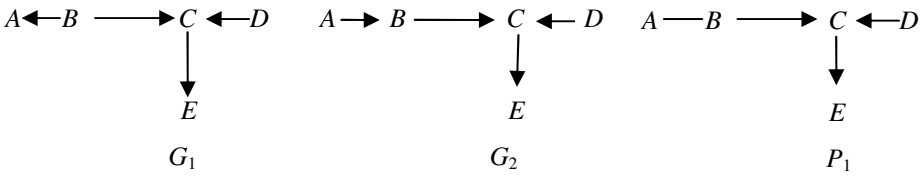


Figure 22: G_1, G_2, **and** P_1 **(pattern for** G_1 **and** G_2**)**

A probability distribution $P(\mathbf{V})$ satisfies the *local Markov condition* for a DAG G_1 with vertices **V** when each variable is independent of its non-parental non-descendants conditional on its parents. A *Bayesian network* is an ordered pair of a directed acyclic graph *G* and a set of probability distributions that satisfy the local Markov condition for *G*.

The graphical relationship among sets of variables in a DAG *G* called "d-separation" determines which conditional independence relations are entailed by satisfying the local directed Markov property). Following [Pearl 1988], in a DAG *G*, for disjoint variable sets **X**, **Y**, and **Z**, **X** and **Y** are *d-separated* conditional on **Z** in *G* if and only if there exists no path *U* between an $X \in \mathbf{X}$ and a $Y \in \mathbf{Y}$ such that (i) every collider on *U* has a descendent in **Z** and (ii) no other vertex on *U* is in **Z**. An important theorem in [Pearl 1988] is that a DAG *G* entails that **X** is independent of **Y** conditional on **Z** if and only if **X** is d-separated from **Y** conditional on **Z** in *G*.

A Bayesian network restricted to a parametric family <*G*,**Θ**> where *G* is a DAG and **Θ** is some parameterization of the DAG, e.g. multivariate Gaussian, has two distinct interpretations. First, it has a probabilistic interpretation as a distribution over the variables in *G*, for distributions that satisfy the local Markov condition for *G*. Under this interpretation, it is a useful tool for calculating conditional probabilities.

Second, it has a causal interpretation, and can be used to calculate the effects of manipulations. Intuitively, a manipulation of a variable is an exogenous action that *forces* a value (or a distribution over values) upon a variable in the system, e.g. as in a randomized experiment - if no exogenous action is taken on variable *X*, *X* is said to have undergone a null manipulation. An example of a manipulation is a randomized experiment, in which a distribution for some variables (e.g. Ω of the subjects take a given drug, and Ω of the subjects do not take the given drug) is imposed from outside. The kinds of manipulations that we will consider are ideal in the sense that a manipulation of *X* directly affects only *X*.

X is a *direct cause* of *Y* relative to a set of variables **V** if there is a pair of manipulations (including possibly null manipulations, and including hypothetical

manipulations in the many circumstances where no actual manipulations are feasible) of the values of the variables in $\mathbf{V}\backslash\{Y\}$ that differ only in the value assigned to X, but that have different distributions for Y. This is in accord with the idea that the gold standard for determining causation is randomized experiments. (This is not a reduction of causality to non-causal concepts, because manipulation is itself a causal concept that we have taken as primitive.) Under the causal interpretation of DAGs, there is an edge $X \rightarrow Y$ when X is a direct cause of Y relative to the set of variables in the DAG. A set of variables $\mathbf{V}$ is *causally sufficient* if every direct cause (relative to $\mathbf{V}$) of any pair of variables in $\mathbf{V}$, is also in $\mathbf{V}$. We will assume that causally interpreted DAGs are causally sufficient, although we will not generally all of the variables in a causally interpreted DAG are measured.

In automated causal search, the goal is to discover as much as possible about the true causal graph for a population from a sample from the joint probability distribution over the population, together with background knowledge (e.g. parametric assumptions, time order, etc.) This requires having some assumptions that link (samples from) probability distributions on the one hand, and causal graphs on the other hand. Extensive discussions of the following assumptions that we will make, including arguments for making the assumptions as well as limitations of the assumptions can be found in *Causation*, *Prediction*, & *Search* [Spirtes et al. 2001].

7.1 Causal Markov Assumption

The Causal Markov Assumption is a generalization of two commonly made assumptions: the immediate past screens off the present from the more distant past; and if X does not cause Y and Y does not cause X, then X and Y are independent conditional on their common causes. It presupposes that while the random variables of a unit in the population may causally interact, the units themselves are not causally interacting with each other.

Causal Markov Assumption: Let G be a causal graph with causally sufficient vertex set $\mathbf{V}$ and let P be a probability distribution over the vertices in $\mathbf{V}$ generated by the causal structure represented by G. G and P satisfy the Causal Markov Condition if and only if for every W in $\mathbf{V}$, W is independent of its non-parental non-descendants conditional on its parents in G.

In graphical terms, the Causal Markov Assumption states that in the population distribution over a causally sufficient set of variables, each variable is independent of its non-descendants and non-parents, conditional on its parents in the true causal graph.

While the Causal Markov Assumption allows for some causal conclusions from sample data, it only supports inferences that some causal connections exist - it does not support inferences that some causal connections do not exist. The following assumption does support the latter kind of inference.

7.2 Causal Faithfulness Assumption

Often the set of distributions that satisfy the local Markov condition for G is restricted to some parametric family Θ (e.g. Gaussian). In those cases, the set of

distributions belonging to the Bayesian network will be denoted as $\boldsymbol{f}(<G, \boldsymbol{\Theta}>)$, and $f(<G,\theta>)$ will denote a member of $\boldsymbol{f}(<G, \boldsymbol{\Theta}>)$ for the particular value $\theta \in \boldsymbol{\Theta}$ (and $f(<G,\theta>)$ is **represented** by $<G,\boldsymbol{\Theta}>$). Let $I_f(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ denote that **X** is independent of **Y** conditional on **Z** in a distribution *f*.

If a DAG *G* does not entail that $I_{f(G,\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$ for *all* $\theta \in \boldsymbol{\Theta}$, nevertheless there may be *some* parameter values θ such that $I_{f(G,\theta)}(\mathbf{X},\mathbf{Y}|\mathbf{Z})$. In that case say that $f(<G,\theta>)$ is **unfaithful** to *G*. In Pearl's terminology the distribution is *unstable* [Pearl 1988]. This would happen for example if taking birth control pills increased the probability of blood clots directly, but decreased the probability of pregnancy which in turn increased the probability of blood clots, and the two causal paths exactly cancelled each other. We will assume that such unfaithful distributions do not happen - that is there may be such canceling causal paths, but the causal paths do not exactly cancel each other.

Causal Faithfulness Assumption: For a true causal graph *G* over a causally sufficient set of variables **V**, and probability distribution $P(\mathbf{V})$ generated by the causal structure represented by *G*, if *G* does not entail that **X** is independent of **Y** conditional on **Z** then **X** is not independent of **Y** conditional on **Z** in $P(\mathbf{V})$.

7.3 Conditional Independence Equivalence

Let $\mathbf{I}(<G,\boldsymbol{\Theta}>)$ be the set of all conditional independence relations entailed by satisfying the local Markov condition. For any distribution that satisfies the local directed Markov property for *G*, all of the conditional independence relations in $\mathbf{I}(<G,\boldsymbol{\Theta}>)$ hold. Since these independence relations don't depend upon the particular parameterization but only on the graphical structure and the local directed Markov property, they will henceforth be denoted by $\mathbf{I}(G)$.

G_1 and G_2 are *conditional independence equivalent* if and only if $\mathbf{I}(G_1) = \mathbf{I}(G_2)$. This occurs if ad only if G_1 and G_2 have the same d-separation relations. A set of graphs that are all conditional independence equivalent to each other is a *conditional independence equivalence class*. If the graphs are all restricted to be DAGs, then they form a *DAG conditional independence equivalence class*.

Theorem 1 (Pearl, 1988): Two directed acyclic graphs are conditional independence equivalent if and only if they contain the same vertices, the same adjacencies, and the same unshielded colliders.

For example, Theorem 1 entails that the set consisting of G_1 and G_2 in

Figure 22 is a DAG conditional independence equivalence class. The fact that G_1 and G_2 are conditional independence equivalent, but are different causal models, indicates that in general any algorithm that relies only on conditional independence relations to discover the causal graph cannot (without stronger assumptions or more background knowledge) reliably output a single DAG. A reliable algorithm could at best output the DAG conditional independence equivalence class, e.g. $\{G_1, G_2\}$.

Fortunately, Theorem 1 is also the basis of a simple representation called a pattern [Verma & Pearl 1990] of a DAG conditional independence equivalence class. Patterns

can be used to determine which predicted effects of a manipulation are the same in every member of a DAG conditional independence equivalence class and which are not.

The adjacency phase of the PC algorithm is based on the following two theorems, where **Parents**(G,A) is the set of parents of A in G.

Theorem 2: If A and B are d-separated conditional on any subset **Z** in DAG G, then A and B are not adjacent in G.

Theorem 3: A and B are not adjacent in DAG G if and only if A and B are d-separated conditional on **Parents**(G,A) or **Parents**(G,B) in G.

The justification of the orientation phase of the PC algorithm is based on Theorem 4.

Theorem 4: If in a DAG G, A and B are adjacent, B and C are adjacent, but A and C are not adjacent, either B is in every subset of variables **Z** such that A and C are d-separated conditional on **Z**, in which case <A,B,C> is not a collider, or B is in no subset of variables **Z** such A and C are d-separated conditional on Z, in which case <A,B,C> is a collider.

A **pattern** (also known as a PDAG) P represents a DAG conditional independence equivalence class **X** if and only if:

1. P contains the same adjacencies as each of the DAGs in **X**;
2. each edge in P is oriented as $X \rightarrow Z$ if and only if the edge is oriented as $X \rightarrow Z$ in every DAG in **X**, and as X — Z otherwise.

There are simple algorithms for generating patterns from a DAG [;1995 ,Meek 1995Chickering ;1997Perlman & ,Madigan ,Andersson**]**. The pattern P_1 for the DAG conditional independence equivalence class containing G_1 is shown in

Figure **22**. It contains the same adjacencies as G_1, and the edges are the same except that the edge between A and B is undirected in the pattern, because it is oriented as $A \leftarrow B$ in G_1, and oriented as $A \rightarrow B$ in G_2.

7.4 Distributional Equivalence

For multi-variate Gaussian distributions and for multinomial distributions, every distribution that satisfies the set of conditional independence relations in **I**(<G,Θ>) is also a member of $\boldsymbol{f}$(<G,Θ>). However, for other families of distributions, it is possible that there are distributions that satisfy the conditional independence relations in **I**(<G,Θ_α>), but are not in $\boldsymbol{f}$(<G,Θ_α>), i.e. the parameterization imposes constraints that are not conditional independence constraints [Lauritzen et al. 1990; Pearl 2000; Spirtes et al. 2001].

It can be shown that when restricted to multivariate Gaussian distributions, G_1 and G_2 in

Figure 22 represent exactly the same set of probability distributions, i.e. $\boldsymbol{f}$(<G_1,Θ_1>) = $\boldsymbol{f}$(<G_2,Θ_2>). In that case say that <G_1,Θ_1> and <G_2,Θ_2> are *distributionally equivalent*

(relative to the parametric family). Whether two models are distributionally equivalent depends not only on the graphs in the models, but also on the parameterization families of the models. A set of models that are all distributionally equivalent to each other is a *distributional equivalence class*. If the graphs are all restricted to be DAGs, then they form a *DAG distributional equivalence class*.

In contrast to conditional independence equivalence, distribution equivalence depends upon the parameterization families as well as the graphs. Conditional independence equivalence of G_1 and G_2 is a necessary, but not always sufficient condition for the distributional equivalence of $<G_1,\mathbf{\Theta_A}>$ and $<G_2,\mathbf{\Theta_B}>$.

Algorithms that rely on constraints beyond conditional independence may be able to output subsets of conditional independence equivalence classes, although without further background knowledge or stronger assumptions they could at best reliably output a DAG distribution equivalence class. In general, it would be preferable to take advantage of the non conditional independence constraints to output a subset of the conditional independence equivalence class, rather than simply outputting the conditional independence equivalence class. For some parametric families it is known how to take advantage of the non conditional independence constraints (sections 3.4 and 4.4); however in other parametric families, either there are no non conditional independence constraints, or it is not known how to take advantage of the non conditional independence constraints.

8 References

Ali, A. R., Richardson, T. S., & Spirtes, P. (2009). Markov Equivalence for Ancestral Graphs. *Annals of Statistics*, *37*(*5B*), 2808-2837.

Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. *Proceedings of the 2003 American Medical Informatics Association Annual Symposium*. (pp. 21-25). Washington, DC.

Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann Stat*, *25*(2), 505-541.

Asuncion, A. & Newman, D. J. (2007). UCI Machine Learning Repository.

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists (Texts in Statistical Science Series).* Chapman & Hall/CRC.

Cartwright, N. (1994). *Nature's Capacities and Their Measurements (Clarendon Paperbacks).* Oxford University Press, USA.

Chickering, D. M. (2002). Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, *3*, 507-554.

Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*.

Chu, T., & Glymour, C. (2008). Search for Additive Nonlinear Time Series Causal Models. *Journal of Machine Learning Research*, 9(May):967--991, 2008.

Dempster, A. (1972). Covariance selection. *Biometrics*, *28*, 157-175.

Fukumizu, K., Gretton, A., Sun, X., & Scholkopf, B. (2008). *Kernel Measures of Conditional Dependence. Advances in Neural Information Processing Systems 20*

Geiger, D. & Meek, C. (1999). *Quantifier Elimination for Statistical Problems*. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden.

Gierl, M. J. & Todd, R. W. (1996). A Confirmatory Factor Analysis of the Test Anxiety Inventory Using Canadian High School Students. *Educational and Psychological Measurement*, *56*(2), 315-324.

Glymour, C. (1998). What Went Wrong? Reflections on Science by Observation and the Bell Curve. *Philosophy of Science*, *65*(1), 1-32.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling.* Academic Press.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Scholkopf, B., & Smola, A. J. A kernel statistical test of independence, In *Advances in Neural Information Processing Systems 20.*

Harman, H. H. (1976). *Modern Factor Analysis.* University Of Chicago Press.

Hoover, K. & Demiralp, S. (2003). Searching for the Causal Structure of a Vector Autoregression. *Oxford Bulletin of Economics and Statistics 65 (Supplement)*, *65*, 745-767.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Scholkopf, B. (2009). *Nonlinear causal discovery with additive noise models. Advances in Neural Information Processing Systems 21*

Hoyer, P. O., Shimizu, S., & Kerminen, A. (2006). *Estimation of linear, non-gaussian causal models in the presence of confounding latent variables*. Paper presented at the Third European Workshop on Probabilistic Graphical Models.

Hoyer, P. O., Hyvärinen, A., Scheines, R., Spirtes, P., Ramsey, J., Lacerda, G. et al. (2008). *Causal discovery of linear acyclic models with arbitrary distributions*. *Proceedings of the Twentyfourth Annual Conference on Uncertainty in Artificial Intelligence*.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5): 411 - 430.

Junning, L. & Wang, Z. (2009). Controlling the False Discovery Rate of the Association/Causality Structure Learned with the PC Algorithm1. *Journal of Machine Learning Research,* 475 - 514.

Kalisch, M. & Buhlmann, P. (2007). Estimating high dimensional directed acyclic graphs with the PC algorithm. *Journal of Machine Learning Research*, *8*, 613-636.

Kiiveri, H. & Speed, T. (1982). Structural analysis of multivariate data: A review. In S. Leinhardt (Ed.), *Sociological Methodology 1982.* San Francisco: Jossey-Bass.

Klepper, S. (1988). Regressor Diagnostics for the Classical Errors-in-Variables Model. *J Econometrics*, *37*(2), 225-250.

Klepper, S., Kamlet, M., & Frank, R. (1993). Regressor Diagnostics for the Errors-in-Variables Model - An Application to the Health Effects of Pollution. *J Environ Econ Manag*, *24*(3), 190-211.

Lacerda, G., Spirtes, P., Ramsey, J., & Hoyer, P. O. (2008). *Discovering Cyclic Causal Models by Independent Component Analysis*. *Proceedings of the 24th Conference on Uncertainty In Artificial Intelligence*.

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

Linthurst, R. A. (1979). *Aeration, nitrogen, pH and salinity as factors affecting Spartina Alterniflora growth and dieback.*

Meek, C. (1995). Causal inference and causal explanation with background knowledge. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence.*

Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence.*

Moneta, A. & Spirtes, P. (2006). *Graphical Models for the Identification of Causal Structures in Multivariate Time Series Model*. Paper presented at the *2006 Joint Conference on Information Sciences*.

Needleman, H. L. (1979). Lead levels and children's psychologic performance. *N Engl J Med*, *301(3)*(3), 163.

Needleman, H. L., Geiger, S. K., & Frank, R. (1985). Lead and IQ scores: a reanalysis. *Science*, *227(4688)*(4688), 701-2, 704.

Pearl, J. & Dechter, R. (1996). Identifying independencies in causal graphs with feedback. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, Portland, OR.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Rawlings, J. (1988). *Applied Regression Analysis.* Belmont, CA: Wadsworth.

Richardson, T. S. (1996). *A discovery algorithm for directed cyclic graphs*. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, Portland, OR.

Scheines, R. (2000). Estimating Latent Causal Influences: TETRAD III Variable Selection and Bayesian Parameter Estimation: the effect of Lead on IQ. In P. Hayes (Ed.), *Handbook of Data Mining.* Oxford University Press.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461-464.

Sewell, W. H. & Shah, V. P. (1968). Social Class, Parental Encouragement, and Educational Aspirations. *Am J Sociol*, *73*(5), 559-572.

Shimizu, S., Hoyer, P. O., & Hyvarinen, A. (2009). Estimation of linear non-Gaussian acyclic models for latent factors. *Neurocomputing*, *72*(7-9), 2024-2027.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, *7*, 2003-2030.

Shpitser, I. & Pearl, J. (2008). Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, *9*, 1941-1979.

Silva, R. & Scheines, R. (2004). Generalized Measurement Models. *reports-archive.adm.cs.cmu.edu*.

Silva, R., Scheines, R., Glymour, C., & Spirtes, P. (2006). Learning the structure of linear latent variable models. *J Mach Learn Res*, *7*, 191-246.

Spearman, C. (1904). General Intelligence objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Spirtes, P. (1995). *Directed cyclic graphical representations of feedback models*. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada.

Spirtes, P. & Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review*, *9*(1), 67-72.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search* (2nd ed.), The MIT Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search.* Spring-Verlag Lectures in Statistics.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, Prediction, and Search, Second Edition (Adaptive Computation and Machine Learning).* The MIT Press.

Spirtes, P., Scheines, R., Glymour, C., & Meek, C. (2004). Causal Inference. In D. Kaplan (Ed.), *SAGE Handbook of Quantitative Methodology*. (pp. 447-477). SAGE Publications.

Strotz, R. H. & Wold, H. O. A. (1960). Recursive VS Nonrecursive Systems- An Attempt At Synthesis. *Econometrica*, *28*(2), 417-427.

Sun, X. (2008). *Causal inference from statistical data.* MPI for Biological Cybernetics.

Swanson, N. R. & Granger, C. W. J. (1997). Impulse Response Function Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions. *Journal of the American Statistical Association*, *92*(437), 357-367.

Tillman, R. E., Danks, D., & Glymour, C. (2009). *Integrating Locally Learned Causal Structures with Overlapping Variables*. In *Advances in Neural Information Processing Systems 21*.

Tillman, R. E., Gretton, A., & Spirtes, P. (2009). *Nonlinear directed acyclic structure learning with weakly additive noise models*. In *Advances in Neural Information Processing Systems 22*.

Timberlake, M. & Williams, K. R. (1984). Dependence, Political Exclusion And Government Repression - Some Cross-National Evidence. *Am Sociol Rev*, *49*(1), 141-146.

Verma, T. S. & Pearl, J. (1990). Equivalence and Synthesis of Causal Models. In P. Bonissone, M. Henrion, L. Kanal, & J. Lemmer (Eds.), *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*. (pp. 220-227).

Voortman, M. & Druzdzel, M. (2008). *Insensitivity of Constraint-Based Causal Discovery Algorithms to Violations of the Assumption of Multivariate Normality*. Paper presented at the FLAIRS Conference 2008.

Zhang, K. & Hyvarinen, A. (2009). *On the Identifiability of the Post-Nonlinear Causal Model. Proceedings of the 26th International Conference on Uncertainty in Artificial Intelligence.*