Available Technology for Discovering Causal Models, Building Bayes Nets, and Selecting Predictors: The TETRAD II Program

Clark Glymour
Department of Philosophy
Carnegie Mellon University
Pittsburgh PA. 15215
cg09@andrew.cmu.edu

Abstract

This paper describes the facilities available for knowledge discovery in databases using the TETRAD II program. While a year or two shy of state of the most advanced research on discovery, we believe this program provides the most flexible and reliable suite of procedures so far available commercially for discovering causal structure, semi-automatically constructing Bayes networks, estimating parameters in such networks, and updating. The program can also be used to rediuce the number of variables needed for classification or prediction, for example as a neural net pre-proceesor. The theoretical principles on which the program is based are described in detail in Spirtes, Glymour and Scheines (1993). Under assumptions described there, each of the search and discovery procedures we will describe have been proved to give correct information when statistical decisions are made correctly.[1]

1. What Does TETRAD Do? This paper describes the facilities available for knowledge discovery in databases using the TETRAD II program. While a year or two shy of state of the most advanced research on discovery, we believe this program provides the most flexible and reliable suite of procedures so far available commercially for discovering causal structure, semi-automatically constructing Bayes networks, estimating parameters in such networks, and updating. The theoretical principles on which the program is based are described in detail in Spirtes, Glymour and Scheines (1993). Under assumptions described there, each of the search and discovery procedures we will describe have been proved to give correct information when statistical decisions about independence and conditional independence have correct outcomes correct in the population distribution. Each of the procedures has also been extensively tested on simulated data samples of realistic sizes .The program includes:

- A module (BUILD) that combine the user's knowledge about the system under study with principles for extracting causal structure from statistical patterns for data sets with continuous variables, or for data sets with discrete variables. The procedure contains a switch that permits the user to assume, or not, that no latent variables are present.
- Functions that indicate when two or more measured variables may all be influenced by an unmeasured common cause.
- A module (ESTIMATE) that gives maximum likelihood estimates for parameters in statistical models describing influences between measured discrete variables. With discrete data and a little help from the user, BUILD and ESTIMATE will construct a fully parameterized Bayes network for a domain.
- A module (UPDATE) that updates a fully parameterized Bayes network to make predictions about any of the properties of a new unit or example from information about some of the properties in that unit.
- A module (PURIFY) that takes a raw data or a covariance matrix for normally distributed variables the user assumes to have *at least* one unmeasured common cause and finds a subset of variables that have *exactly* one unmeasured common cause and no other causal relations with one another.
- A Module (MIMbuild) that determines structural dependencies among latent variables given correlational data and purified measurement models.
- A module that prepares input files for other estimation and testing packages (EQS, LISREL, CALIS) for linear models.

[1] Research supported by the Navy Office of Personnel Research and Development, and the Office of Naval Research, contract N00014-93-1-0568. This work is a collaboration with C. Meek, R. Scheines and P. Spirtes.

- A module (MONTE) that allows the user to generate simulated data for a wide variety of causal models.

The TETRAD II program does not do routine data cleanup task--checks for outliers, variable transformations to approximate normal distributions, etc. Neither does it do model diagnostics of the kind performed by many readily available statistical packages such as MINITAB, SAS, BMDP or SYSTAT. We recommend that where possible checks and adjustments of the data be carried out first by one of these systems prior to a TETRAD II analysis. The TETRAD II program does not estimate the parameters of linear "structural equation models" or provide tests of significance for such models, since these procedures are carried out by a number of commercial packages such as CALIS, LISREL and EQS.

2. Graphical Models. Many statistical models that are given by equations and distribution assumptions can be described more vividly but equally precisely by simple directed graphs. A directed edge X -> Y indicates both that X influences Y and that Y is a function of at least X. For example, suppose we consider a regression model for Y with regressors X1,...,X4. The model might be given by an equation

$$Y = a_0 + a_1X1 + a_2 X2 + a_3X3 + a_4X4 + \varepsilon$$

and a distribution claim: all variables are jointly normally distributed, each variable in the set {X1,...,X4, ε} is independent of the other variables in the set and ε has mean zero. The statistical model has a number of free parameters that must be estimated from the data. They include the numerical values of the coefficients, a_1, a_2, etc., the variance of ε and the means and variances of X1,...,X4. We could equally describe the model by saying that the variables are jointly normal and giving the picture:

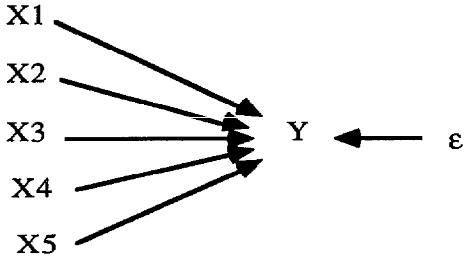

Figure 1

The equations can be easily recovered from the picture by writing each variable as a linear function of the variables with edges directed into it. In this case Y is the only variable with any edge into it. The picture is always understood to imply the hypothesis of statistical independence for any pair of variables, such as X1 and X2, or X3 and ε, that are not connected by a sequence of directed edges from one to the other, or by two sequences of directed edges from some third variable that is a common cause of both. The numerical values of linear coefficients such as a_1 are not specified in the diagram, just as they are not specified in the equation. For brevity "error terms" such as ε are often omitted in the diagram, but they are meant to be added if the diagram is translated into a set of equations. The pictorial representation can be expanded to include unmeasured common causes. In that case we adopt the common convention of writing the measured variables inside rectangles and unmeasured variables (except error terms) inside ovals.

Diagrams can also represent causal and statistical hypotheses among discrete variables. Suppose, for example, that in figure 1 the variables are discrete. Then the statistical hypothesis represented by the diagram is that the joint probability distribution is equal to the conditional distribution of Y on the other six variables, multiplied by the marginal distributions of each of those six. That is:

$$P(Y,X1, X2, X3, X4,X5,\varepsilon) =$$
$$P(Y|X1,X2,X3,X4,X5, \varepsilon) \text{ X}$$
$$P(X1)P(X2)P(X3)P(X4)P(X5)P(\varepsilon)$$

In the computer science literature graphical models that represent a factorization of the probabilities for discrete variables are usually called "Bayes networks."

3. Applications.

Selection of Causal Regressors. Many empirical investigations attempt to judge how much one or more variables influence an outcome of interest, for example, how much advertising influences recruitment or influences purchases of a product. Multiple linear and non-linear regression are the methods most commonly used to make these decisions. The results of a regression can, however, be misleading if some of the regressors are not causes of the outcome variable. A number of techniques are available in commercial packages for selecting a set of regressors from a larger set of variables. Unfortunately, unless the investigator knows beforehand a great deal about the causal relations among the potential regressors, these techniques are unreliable means to determine which variables are actually direct causes of the outcome variable, and regression itself is unreliable in determining the importance of causal influence, no matter how large or representative the sample. TETRAD II can be used to help select direct causes of an outcome variable more reliably.

In data on naval air traffic controller training we analyzed for the Navy Personnel Research and Development Center, trainees were given a battery of tests, and scores on several of these batteries were combined into an "AFQT" score. We obtained data that included the AFQT scores and a battery of test scores, including three test scores--arithmetic reasoning (AR), numerical operations (NO), and word knowledge (WK)--that are components of the AFQT score and four others that are not--electronics information (EI), general science (GS), mechanical comprehension (MC) and mathematical knowledge (MK). AFQT has other measured components that were not included in our data set. A true (but incomplete) account of the dependencies in this data is then:

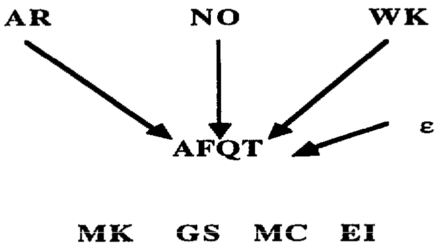

Figure 2

In fact, there is good reason to think there are also unrecorded common causes at work relating these variables with others included in AFQT. For the actual sample data, an ordinary linear regression of AFQT against all of the variables gave significant values for all but two of the coefficients. Given the prior information that the other variables cannot be effects of AFQT, the TETRAD II procedures (BUILD with the exact latent variables switch) determine correctly that among the seven tests, only AR, NO and WK are components of the AFQT. When we first conducted this analysis we had been misinformed that all seven variables are components of AFQT, and the correct answer was thus found without prior knowledge.

Sometimes when we are trying to understand the causes of a variable of interest, or to predict how to manipulate that variable, we may, without knowing it, measure *effects* of the variable we wish to explain and predict. In such cases regression methods may be badly misleading. In many cases of this kind, the TETRAD II procedures can sometimes help. Consider the following diagram:

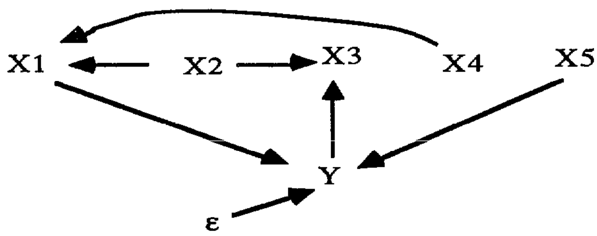

Figure 3

Suppose the task is to find from among X1,...,X5, those variables that actually influence Y directly, and the figure represents the true but unknown structure. If Y is some characteristic we want to change in a population, and the Xs represent variables we think we can manipulate, then we want to find out which of the Xs actually influence Y. We can find linear systems with this causal structure and also systems in which the variables are discrete. In the linear case, for almost any values of the linear coefficients and variances, linear regression will give us false results. Even with population correlations, linear regression will tell us incorrectly that X1, X2, X3, and X5 have significant regression coefficients, and methods for selecting subsets of regressors do no better. Given population data, the Build procedure in TETRAD II will correctly identify X1 and X5 as the only possible causes of Y among the five X variables. For discrete variables logistic regression meets with similar problems. Twenty data sets were generated by C. Meek using Monte Carlo methods from the model in figure 3, each with a sample size of 5,000. The variables and the disturbance with distributed normally and the linear coefficients were generated randomly for each sample. In all 20 data sets a straightforward linear regression inferred that the coefficients for X1, X2, X3, and X5 are significant. Regression selector packages did even worse. In all 20 cases, best subsets (Mallows CP and Adjusted R^2) and stepwise procedures picked either all five X variables or omitted only X4. In all twenty cases, by contrast, the Build procedure in TETRAD II correctly identified only X1, X3, and X5 as adjacent to Y, and for 94% of the adjacencies the procedure found the causal order correctly as well.

The same regression mistakes would occur if in figure 3 X3 and Y were connected by an unmeasured common cause, and in the same way, TETRAD II would give correct information in such a case.

Causal Models for Discrete Variables, or Bayes Networks. Many variables of interest are better measured by categories than by a real variable, and the properties of models of discrete variables have been extensively studied in the last twenty years. Just as with linear models, one of the common uses of models for discrete variables is to attempt to represent and quantify causal dependencies.

The diagram in figure 4 below, called the ALARM network, was developed for use as an emergency medical system (Beinlich, et al. 1989). The variables are all discrete, taking 2, 3 or 4 distinct values. In most instances a directed arrow indicates that one variable is regarded as a cause of another. The physicians who built the network also assigned it a probability distribution: each variable V is given a probability distribution conditional on each vector of values of the variables having edges directed into V.

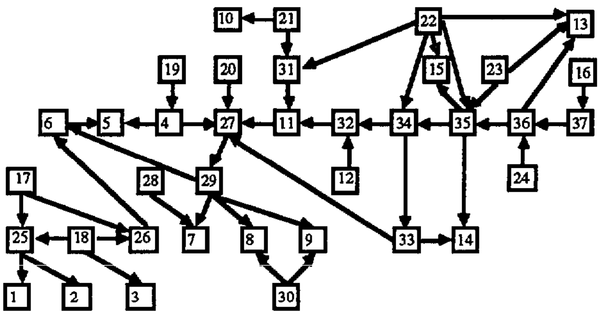

Figure 4

The directed graph has 37 variables and 46 edges. Herskovitz and Cooper (1990) used the diagram to generate simulated emergency medicine statistics for 20,000 individuals. From half or even a tenth of the data, the TETRAD II program recreates almost all of the ALARM network, including information about the directions of the edges. Depending on sample size, the program makes two or three errors in identifying edges and four or five errors in determining the directions of influence. Given the data and a causal diagram, the Estimate module of the TETRAD II program will provide a maximum likelihood estimate (assuming a multinomial distribution) for any Bayes network without cycles or latent variables.

The UPDATE Module. Suppose a Bayes network has been built for a domain, and the probabilities associated with the network have been estimated. The network can function as an "expert system" that will make predictions for new units in the population. Given the measured values of one or more variables for a new unit (for example, a new patient) the program will use the Bayes' network to compute the new probabilities for values of any other variables for that unit. Thus with the ALARM network, if the program is given values for the variables attached to some of the nodes, it will compute a new set of probabilities for the values of any other node in the network.

Finding Causal Relations Involving Unmeasured Variables. Unmeasured variables are important in two roles. On the one hand, they may be responsible for statistical associations among measured variables, and we must then correctly recognize the latent structure if we are to predict the results of policies or interventions. Uniquely, the TETRAD II program contains an asymptotically correct procedure for this problem.

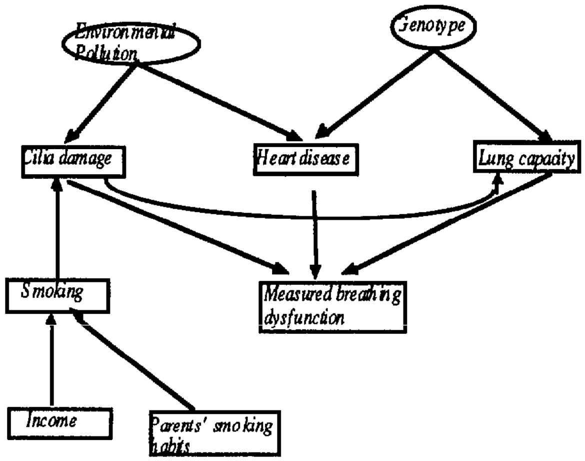

Figure 5

Given large sample data for the variables shown in rectangles but *not* for the variables in ovals in figure 5, the BUILD procedure will reconstruct the graph shown, but with double headed arrows indicating the latent variables, and leaving uncertain whether *income* and *parent's smoking habits* are respectively related to *smoking* by unmeasured common causes.

On the other hand, in many cases the variables we measure are only indicators of the variables that are of interest. That is typically the case in personnel studies, psychometric studies, sociometric studies, and many other cases. Often the data are from questionnaires or other sources for which the investigator has a fairly clear idea as to which sets of measured variables form clusters that indicate the same unmeasured causal factor. The investigator may be fairly confident, for example, that a certain collection of items are all affected by a particular personality factor, that another collection is affected by perceived economic opportunity, and so on. But how are the influences of the unmeasured variables on one another, or on other behavioral variables, to be estimated? Under the assumptions of linearity. normality, and two assumptions characteristic of graphical models (see Spirtes, et al., 1993 for details), when the sample is large enough that correct (in the population distribution) statistical decisions are made, the PURIFY and MIMbuild modules of TETRAD II will provably give reliable information about the connections among the unmeasured variables provided the program is given correct information as to which variables cluster together as indicators of a common unmeasured latent variable. The procedure will work when some of the indicators of one latent variable are also affected by other latent variables, and even when some of the measured variables directly influence one another. Parts of the procedure have been used successfully on real data by Steve Sorenson and his associates at the Navy Personnel Research and Development Center. We will illustrate with a simulated case.

From the linear structure given by the diagram in figure 6 a sample of 2,000 units of the measured variables was generated by the Monte Carlo module of the TETRAD II program. The thicker arrows represent causal relations that confound the original clustering of the measured variables:

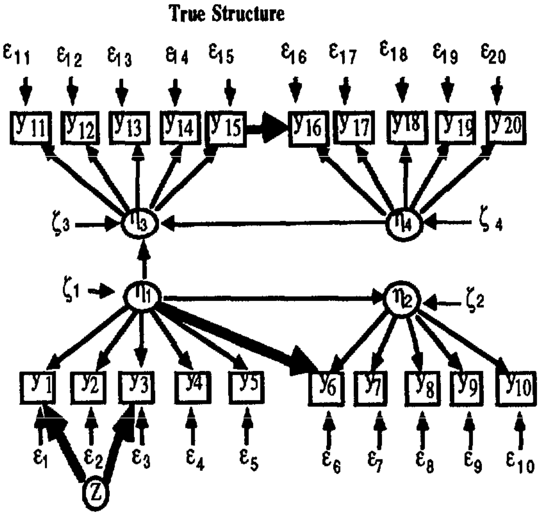
True Structure
ε11 ε12 ε13 ε14 ε15 ε16 ε17 ε18 ε19 ε20
y11 y12 y13 y14 y15 y16 y17 y18 y19 y20
ζ3 η3 η4 ζ4
ζ1 η1 η2 ζ2
y1 y2 y3 y4 y5 y6 y7 y8 y9 y10
ε1 ε2 ε3 ε4 ε5 ε6 ε7 ε8 ε9 ε10
Z

Figure 6

The data for the measured variables was then given to the PURIFY module, along with the clusterings of the measured variables shown in figure 7. For each cluster, TETRAD II automatically finds a sub-collection of measured variables that are not affected by the latent variables of other clusters or by other measured variables. The output of PURIFY is shown in figure 8.

Given the clusters in figure 8 and the original data, the MIMbuild program then returns the correct information about the causal relations among the unmeasured variables, shown in figure 9.

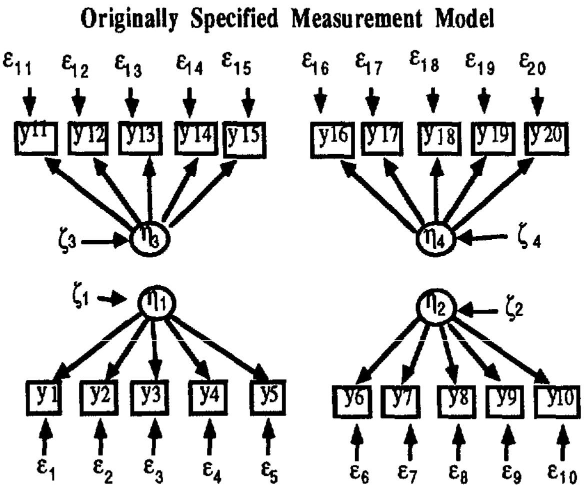
Originally Specified Measurement Model
ε11 ε12 ε13 ε14 ε15 ε16 ε17 ε18 ε19 ε20
y11 y12 y13 y14 y15 y16 y17 y18 y19 y20
ζ3 η3 η4 ζ4
ζ1 η1 η2 ζ2
y1 y2 y3 y4 y5 y6 y7 y8 y9 y10
ε1 ε2 ε3 ε4 ε5 ε6 ε7 ε8 ε9 ε10

Figure 7

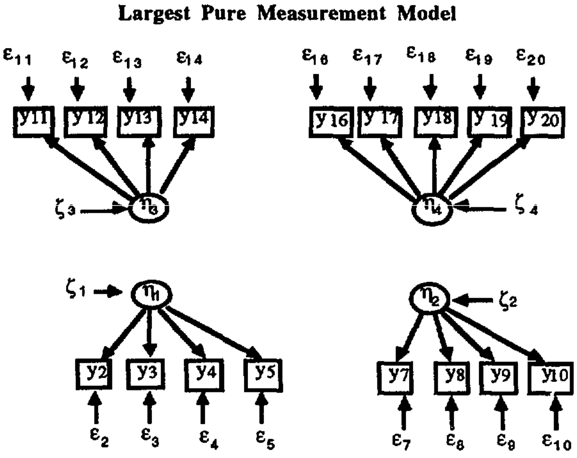
Largest Pure Measurement Model
ε11 ε12 ε13 ε14 ε16 ε17 ε18 ε19 ε20
y11 y12 y13 y14 y16 y17 y18 y19 y20
ζ3 η3 η4 ζ4
ζ1 η1 η2 ζ2
y2 y3 y4 y5 y7 y8 y9 y10
ε2 ε3 ε4 ε5 ε7 ε8 ε9 ε10

Figure 8

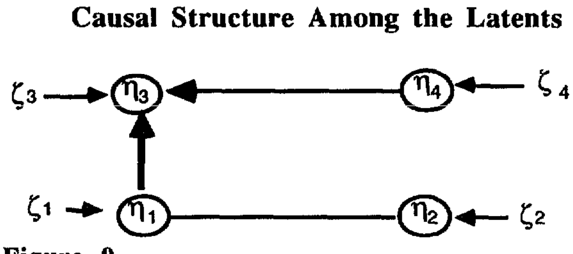

Figure 9

Selection of Prediction Variables. Given a set S of variables and a variable Y one wishes to predict from S, it is often important in applications to minimize the number of variables in S needed for prediction. The Build procedure excludes variables by finding a subset of S condiitonal on which all other variables in S are independent of Y. For example, Neuralware divides Fisher's well known Iris data into a training and test set to illustrate the use of neural nets in a simple classification problem in which variety Y is to be predicted from a set S of four features of flowers. BUILD identifies two of the four variables that suffice for predicting Iris type. When the same Neuralware procedure is applied using only these two variables as predictors, the resulting network misclassifies only a single extra case than does the original larger network, which made a single error in 75 test cases.

Researchers in a number of areas have begun to make use to TETRAD II procedures. The program has been used to develop models of plant metabolism, to study job satisfaction, to develop scales of pain and grief, to detect leakage in frequency channels in devices for measuring sound, to locate structural defects in satellites, to study evoked response potentials, student retention in universities, pneumonia triage, and in other applications.

References:

Beinlich, I., Suermondt, H., Chavez, R., and Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. Proc.Second European Conference on Artificial Intelligence in Medicine, London, England. 247-256.

Cooper, G. and Herskovitz, E. (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning.

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

R. Scheines, P. Spirtes, C. Glymour and C. Meek, *TETRAD II*, Lawrence Erlbaum, 1995.