Regression Can Build Predictive Causal Models

Paul R. Cohen, Lisa A. Ballesteros, Dawn E. Gregory, and Robert St. Amant

Computer Science Technical Report 94-15

Experimental Knowlege Systems Laboratory
Department of Computer Science, Box 34610
Lederle Graduate Research Center
University of Massachusetts
Amherst, MA 01003-4610

Abstract

Covariance information can help an algorithm search for predictive causal models and estimate the strengths of causal relationships. This information should not be discarded after conditional independence constraints are identified, as is usual in contemporary causal induction algorithms. Our FBD algorithm combines covariance information with an effective heuristic to build predictive causal models. We demonstrate that FBD is accurate and efficient. In one experiment we assess FBD's ability to find the best predictors for variables; in another we compare its performance, using many measures, with Pearl and Verma's IC algorithm. And although FBD is based on multiple linear regression, we cite evidence that it performs well on problems that are very difficult for regression algorithms.

This research is supported by ARPA under contract F30602-93-C-0100, and by a NASA GSRP Training Grant, #NGT-70358.

1. Causal Induction

Everyone knows correlation does not imply causation and nobody denies that a strong correlation often suggests a causal relationship. We usually follow up the suggestion with an experiment: if x and y are correlated, and y changes when x is manipulated, then x causes y. But what if we cannot run an experiment? What if we can only *observe* a relationship between x and y? Then it will be difficult to say whether x causes y or vice versa, and we will be unable to say whether something else is causing x and y, both. Thus, some assert that causal inferences without experiments are impossible (e.g.[6]).

As Pearl and Verma point out, though, we cannot spend our lives running controlled experiments, and sometimes we *do* infer cause from associations: "... the question remains how causal knowledge is ever acquired from experience." [11] Pearl and Verma have designed and implemented algorithms to infer causal relationships given only correlational (covariance) information [11, 10]. Spirtes, Glymour, Scheines and their colleagues have developed several algorithms with similar goals and underlying principles [14]. In fact, there have been several efforts in AI (e.g.[7, 4]) dating back to Simon [12] to induce causal relationships from observational data. All rely on *conditional independence* in one way or another. Crudely, if x and y are conditionally independent given z, then either z causes both x and y or z sits between x and y in a causal chain.

Another approach, dating back to Sewall Wright's *path analysis*, is closely related to multiple linear regression [9, 13, 5]. Historically, path analysis was used to *estimate* the strengths of causal influences in a causal model. Someone still had to propose a causal model, but path analysis could provide measures of how well it fit the data. Algorithms based on conditional independence do not estimate strengths of causal relationships. In fact, they use quantitative information only to infer boolean conditional independence constraints, from which they produce causal models that are consistent with the constraints. To estimate strengths of causal influence, these models (and the covariance matrices from which they are derived) are handed off to a statistical package such as LISREL or EQS [8, 2].

We claim that covariance information can guide the search for causal models, estimate the strengths of causal relationships, and yield *predictive* causal models. It shouldn't be thrown away after conditional independence constraints are identified. We have developed an algorithm that builds a predictive model for a dependent variable, then builds predictive models for the predictors, and so on, until it runs out of variables to predict. A simple, intuitive, effective heuristic decides which variables to use as predictors at each level of the model. We will describe the algorithm and the sense in which its models are causal, and we will present the results of two experiments, including an extensive comparison with Pearl and Verma's IC algorithm.

2. Multiple Regression and the FBD Algorithm

Multiple linear regression builds predictive models given a dependent variable (or "predictee") and a set of predictors. A regression model estimates the value of a predictee as a weighted sum of predictors, and the weights are called regression coefficients. Their magnitudes depend on the units of measurement of the predictors; for instance, if a predictor

is age in months, then the regression coefficient will be smaller than if age was measured in years. To compare regression coefficients without fussing about units of measurement, variables are often *standardized* so their means are zero and their standard deviations are one. To standardize a variable, subtract its mean from each value and divide by its standard deviation: $X_i = (x_i - \bar{x})/s$. After standardizing variables, a regression model has this form:

$$\hat{Y}_i = \beta_{X_1} X_{1_i} + \beta_{X_2} X_{2_i} + \cdots + \beta_{X_k} X_{k_i}$$

for standardized predictors $X_1, X_2, \ldots X_k$. The coefficients β_{X_i} are called *beta coefficients* and they have a causal interpretation given shortly.

Multiple regression finds beta coefficients so that the regression model is the best possible in the *least squares* sense of minimizing $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. A set of predictors accounts for some fraction of the variance in the dependent variable. We denote the variance SS_{total}—the summed, squared deviations of values of Y from their mean. If the regression model always made accurate predictions, if $\hat{Y}_i - Y_i$ was always zero, then we'd say the model accounted for 100% of the variance in Y. In general, though, the *residuals* $\hat{Y}_i - Y_i$ are not zero, and the sum of the squared residuals (called SS_{error}) is some fraction of the total variance in Y. Ideally SS_{error} will be very small. The proportion $R^2 = (SS_{total} - SS_{error})/SS_{total}$ is 1.0 if the regression model accounts for all the variance in Y.

Note that regression models are only one *level* deep, in the sense that the predictors point directly to the dependent variable, as opposed, say, to X_2 predicting X_1 which in turn predicts Y. This is unfortunate because regression models have many advantages: they are predictive, least-squares models; and beta coefficients can be compared to each other, and they have a causal interpretation. The FBD algorithm constructs *multi-level* regression models; for example, Figure 1.b is the one it built to fit the data generated by 1.a. Before we describe it in detail, let's consider the sense in which multi-level regression models are causal models.

2.1 Beta Coefficients as Strengths of Causal Influence

If you subscribe to one notion of what a "cause" is, then beta coefficients are strengths of causal influence. Suppes [15] posits three requirements for x to cause y: x and y must *covary*, x must *precede* y, and other causes of y must be *controlled*. For instance, you flip a light switch 100 times and record that on 99 occasions, the light turns on; this is covariance. Moreover, the light never goes on before you flipped the switch; this is precedence. You consider the possibility that you are *willing* the light to go on and switch- flipping is merely incidental, but when you strap your arm to a chair and intone "the light will go on," nothing happens. This is control. The big problem for philosophers, economists, AI researchers, and others, is whether you can infer that flipping the switch caused the light to go on if you *cannot* run a control condition as just described. Some people say yes, some say no, but all acknowledge that *statistical control* can play a similar role to experimental control. One example of statistical control is the partial correlation coefficient, the correlation between two factors when the influences of other factors are held constant. For example, the correlation of shoe size and language ability is very high, but the partial correlation of these factors when age is held constant is roughly zero. Beta coefficients are partial in the same sense.

A beta coefficient β_{X_1} represents the influence of a predictor X_1 on Y when the influences of all other predictors are held constant. And because betas are coefficients of standardized variables, they are in the same units of measurement, so if $\beta_{X_1} = .8$ and $\beta_{X_2} = .4$, then X_1 has twice the influence on Y that X_2 has.

Thus a regression model is causal according to Suppes' criteria: If X and Y don't covary, then $\beta_X = 0$. If X is the predictor and Y the dependent variable, then in the regression model X precedes Y (the model could be wrong, of course; the point is it includes precedence relationships as Suppes requires). Finally, β_X provides a statistical version of control.

2.2 Finding Good Multi-level Models

The FBD algorithm is told which variable is at the root of a causal model (e.g., **var-9** in Figure 1) and it is also given a dataset that includes variates of this variable and potential predictors. It finds a subset of predictor variables that explain the predictee (e.g., it selected **1,5,7** and **8**); then it backs up and treats each of these predictors as a new predictee. Note that the set of predictors doesn't change from one level to another; hence **8** and **5** both predict **9** and also **7**, which itself predicts **9**.

We assess the goodness of predictors as follows: Let Y be the predictee and $X_1, X_2, \ldots X_k$ be predictors. A beta coefficient β_{X_i} measures the strength of influence of X_i on Y when the influences of all other predictors are held constant, and a correlation coefficient r_{YX_i} is the *total* influence of X_i on Y. Thus, $r_{YX_i} - \beta_{X_i}$ is the influence of X_i on Y that is due to X_i's relationships with other predictors, and

$$\omega_i = \frac{r_{Y,X_i} - \beta_{X_i}}{r_{Y,X_i}}$$

is the proportion of X_i's influence on Y that is *not* direct. For example, the correlation between variables **5** and **9** (see Figure 1) is $r_{59} = .58$ and, when **9** is regressed on *all* eight predictors, $\beta_{59} = .53$. Hence, $\omega_{59} = (.58 - .53)/.58 = .086$, which means only nine percent of **5**'s influence on **9** is due to its relationships with other predictors: **5** is a good predictor. The FBD algorithm selects predictors that have ω scores less than a threshold, T_ω. However, it sometimes happens that both r_{Y,X_i} and β_{X_i} are very small, and ω_i is, too, even though X_i has almost no influence on Y. Thus, if β_{X_i} is less than a threshold T_β, we discard X_i. Predictors that survive these two tests are subjected to two more: A set of predictors must account for at least some of the variance in the predictee, so we require R^2 for the regression of the predictor on the set to exceed a threshold, T_{R^2}. Finally, each member of the set of predictors should not be conditionally independent of the predictee, given the other members.[1] Because this conditional independence test did not affect FBD's behavior we will discuss it no further here, and present the algorithm without it.

[1]We found, however, that the sets of predictors selected by ω scores were always consistent—no member was rendered conditionally independent of the predictee by the other members. A *single* member often rendered another conditionally independent of the predictee, but never was a member made conditionally independent by *all* the other members. Possible explanations for this phenomenon, and their consequences for the control of causal induction algorithms, are discussed in [3].

3. The FBD Algorithm

Briefly, here is the FBD algorithm. Let ω_{XY} and β_{XY} denote X's ω and β scores when X is used as a predictor of Y, and let T_ω, T_β and T_{R_2} be thresholds as described above. Set *predictees* $= Y$ and *predictors* $= X_1, X_2, \ldots, X_k$

For each predictee y$\in$ **predictees**
  Find a set $p \subset$ *predictors* such that
    $R_p^2 > T_{R^2}$
    for every $x \in p$:
      a link from x to y would not create a cycle
      $\omega_{xy} < T_\omega$
      $\beta_{xy} > T_\beta$
  Add the elements of p to **predictees**
  Remove y from **predictees**
If **predictees** is empty, stop

Note that this algorithm runs *two* regressions for each predictee: The first—with all potential predictors—yields beta coefficients that are used to calculate ω scores and select good predictors. The second— run with the set of surviving good predictors—yields beta coefficients that are interpreted as strengths of causal influence. Typically, a coefficient is higher in the second regression than in the first, but this is not always so, and occasionally it drops below T_β. In the current version of the algorithm, we keep the predictor anyway. This is why the link from 4 to 8 in Figure 1 has a value of .066, even though T_β was set at .1 for the trial.

The FBD algorithm is quite fast. Because it must perform a multiple regression on each of the variables encountered, its time complexity depends on the calculation of the regression coefficients. This is $O(n^3)$, giving a time complexity of $O(n^4)$ for the algorithm. In practice, it takes less than 15 seconds to build models of 6 to 9 variables (100 variates per variable) on a Sparc 10.

4. Experiments with the FBD Algorithm

The experiments reported here used a set of 60 *target models* (see [3] for experiments with a more extensive set). Models included 6, 9 or 12 variables. A target model is a directed acyclic graph with weights on its links and a set of 100 data that are probably sufficient to recreate the model given only the data. Datasets are only "probably" sufficient to recreate models because the data are generated by random sampling, and it's always possible that a dataset will fit some *other* model better than it fits the one it is supposed to fit (see sec 4.2.3). The procedure for generating models is described in [3] and is similar to the procedure in [14].

4.1 Experiment 1. Does FBD Choose Good Predictors?

Because FBD works backwards—first predicting the dependent variable, then predicting the predictors—its ability to choose good predictors is crucial. Let R_P^2 be the proportion of

variance in Y accounted for by a regression model that includes all the predictors in P. Let p and q be disjoint subsets of P. In general, $R^2_p > R^2_q$, but this does not mean we should include elements of q as predictors: they might contribute very little to R^2_p and would serve more usefully as predictors of one or more elements of p. Imagine we want to select the best three predictors for Y, that is, a set u such that $R^2_u > R^2_v$ for any sets $v \neq u$ of size three. We could find the set u by exhaustive search, but this will usually be enormously expensive. Instead, the FBD algorithm uses ω scores to rank predictors, selecting those with $\omega < T_\omega$. Imagine the algorithm selects three predictors for inclusion in p. Are these the best predictors, or is there another set of three such that $R^2_u > R^2_p$? The following experiment suggests that the algorithm selects sets that are nearly as good as the best sets.

We asked how well ω scores selected predictors for the dependent variable in each of our 60 models. Call the dependent variable Y and, for a 12 variable model, the predictors are $P = \{X_1, \ldots, X_{11}\}$. We find the best k predictors of Y by running regressions of Y on all possible subsets of size k of P. The best subset, denoted *p–best*(k), is the one with the highest R^2. Next we select a subset of k predictors using ω scores. The *batch discarding* method regresses Y on $X_1, \ldots, X_{11}$, calculates ω for each of these predictors, and selects the k predictors with the best ω scores. This set is denoted *p–batch*(k). The *iterative discarding* method repeatedly regresses Y on a set of predictors, discarding the predictor with the worst ω score, until only k predictors remain, denoted *p–iter.*(k). If ω scores select good predictors, then *p–best*(k), *p–batch*(k), and *p–iter.*(k) should contain the same predictors, and if they don't, then *p–batch*(k) and *p–iter.*(k) should account for nearly as much of the variance in Y as *p–best*(k).

Table 1 shows how many predictors *p–batch*(k) and *p–iter.*(k) have in common with *p–best*(k). For 12-variable models and $k = 5$, the mean number of predictors shared by *p–batch*(k) and *p–best*(k) is 3.15, and, for *p–iter.*(k) and *p–best*(k) this number is 3.375. Thus, when batch discarding selects five variables, roughly two of them (on average) are not the best variables to select. On the other hand, Table 2 shows that the variables in *p–batch*(k) and *p–iter.*(k) account for almost as much of the variance in Y as those in *p–best*(k). Let $R^2_{p-best(k)}$ be the variance in Y that is "available" to predictors in *p–batch*(k) and *p–iter.*(k). For instance, if the best k predictors account for only 50% of the variance in Y, then we want to express the predictive power of *p–batch*(k) as a fraction of 50%, not 100%. Table 2 therefore contains the ratios $R^2_{p-batch(k)}/R^2_{p-best(k)}$ and $R^2_{p-iter.(k)}/R^2_{p-best(k)}$. For example, for 12-variable models and $k = 5$, the predictors selected by batch discarding account for 85% of the available variance in Y, on average, and those selected by iterative discarding account for 86%.

Batch and iterative discarding do not find exactly the same predictors as exhaustive search for the best predictors, but the ones they find account for much of the available variance in the dependent variable. Bear in mind that exhaustive search for the best k of N variables requires N-choose-k multiple regressions with k predictors, whereas iterative discarding requires $N - k$ multiple regressions with between N and $k + 1$ predictors, and batch discarding requires just one regression with N predictors. Thus, batch discarding finds predictors that are nearly as good as exhaustive search, with a fraction of the effort.

We wondered whether something simpler than ω scores, such as sorting by beta coefficients, would perform as well. We discovered that in most cases, beta coefficients selected the same predictors as ω scores, but sometimes they recommended different sets with bad

R^2 scores. We can see why (and also why ω is preferable) by considering four cases:

High r_{xy}, high β_{XY}: In this case ω is small in absolute value and X will be accepted as a predictor. Similarly, β is high, so by this criterion X will also be accepted as a predictor. Since X should be accepted in this case, both ω and β do the right thing.

High r_{xy}, low β_{XY}: Here, ω is large in absolute value and β small, so both statistics will reject X, which is the right thing to do.

Low r_{xy}, high β_{XY}: Here, ω is large in absolute value so X will be rejected. However, its β score suggests accepting it. The correct action is to reject X because the only way to get, say, $r_{xy} = 0$ and $\beta_{XY} = .8$ is for X's direct influence (.8) to be cancelled by its indirect influence (-.8) through other predictors. We want predictors that have large direct influence and small indirect influence, so we ought to discard X. In this case, β coefficients make the wrong recommendation.

Low r_{xy}, low β_{XY}: In this case, ω is small but FBD will discard X as a predictor because β is small.

4.2 Experiment 2: Performance of FBD on Target Models

We now consider the quality of the models FBD builds from datasets. We assessed eighteen different measures of its performance, summarized in Table 4 and, for most of these measures, compared FBD's performance with that of Pearl and Verma's IC algorithm [11]. The algorithms differ in three important ways. First, FBD is told which variable is the root of each target model, whereas the IC algorithm is given no such information.[2] Second, FBD runs regressions to find good predictors whereas the IC algorithm works with conditional independence relationships, only. Third, FBD models contain only directed links, and they all have the causal interpretation given in section 2.1; whereas models produced by the IC algorithm contain directed links that are considered "genuine causes," as well as other directed links and undirected links. Many of the measures in Table 4 make sense only for directed links, so, unless otherwise noted, undirected links found by the IC algorithm are ignored for the purpose of calculating these measures.

We ran FBD and the IC algorithm on the test set of 60 models described in section 4.. The procedure for FBD was straightforward: we told it the dependent variable for each model and gave it the datasets for the target models and scored the models it returned. FBD's parameters (section 3.) were $T_\omega = 1.0$, $T_\beta = 0.1$, $T_{R^2} = 0.1$, and we used batch discarding of predictors (section 3.). The IC algorithm takes as input a covariance matrix for all the variables in a dataset and, from this, it infers conditional independence relationships. We tried several settings of its parameters, obtaining the best results (reported here) with a separating set of size one and a partial threshold of .1.

The measures in Table 4 will be easier to understand in the context of Figure 1. For brevity we will refer to the target model (Fig.1.a) as TARGET-0, FBD's model (Fig.1.b) as

[2]Thus, IC solves a more difficult problem than FBD. We are not claiming one algorithm is "better" than the other. We include results for the IC algorithm only so we can compare ours to another, established algorithm.

FBD-0, and IC's model (Fig.1.c) as IC-0. The latter two models are referred to as *inferred* models.

4.2.1 Dependent R^2.

The dependent variable in TARGET-0 is **9** and it has five predictors. FBD found four of them. R^2 for the dependent variable for FBD-0 is .625. This is almost identical to R^2 for TARGET-0, .628, which suggests that the fifth variable—the one FBD missed—explains virtually none of the variance in **9**. In fact, looking at TARGET-0, we see that $\beta_4 = -.058$, confirming our suspicion that it contributes little. FBD dropped it because $\beta_4 < T_\beta$. For IC-0, R^2 is .336; relatively low because we require a directed link between variables to consider one a predictor of another. (If nodes connected to the dependent variable by undirected links are also counted, then R^2 is .582 for IC-0). Table 4 shows that FBD's average scores for Dependent R^2 hover around .75, and IC's scores range from .3 to .45, depending on the model size. (If we count undirected links, then IC's average scores range from .6 to .81.)

4.2.2 Nopenalty ΔR^2.

We summarized FBD's ability to predict *all* the variables that it should as follows: Let **V** be a variable in the target model that has other variables pointing to it; calculate R^2 for **V** in the target model and R^2 for **V** in FBD's model. Take the absolute value of the difference of these scores. Then take the mean absolute difference over all such variables in the target model. For example, in TARGET-0, variables **9**, **8** and **6** should be predicted, so the Nopenalty ΔR^2 score is the mean of three absolute differences in R^2 scores. This statistic assesses no penalty for nodes that *shouldn't* be predicted but are, such as **1** and **7** in FBD-0. On average, the mean absolute difference between the target model R^2s and FBD R^2s ranges from .13 (for 6-variable models) to .19 (for 12-variable models). For the IC algorithm the mean absolute differences are all roughly .35. If undirected links are allowed (see above) then these differences are smaller. Because this statistic measures mean *absolute* differences, it could be high if the target R^2s are higher than the inferred model R^2s, or vice versa. In practice, the former was far more common. Thus we can interpret this statistic as saying that FBD models account for 13% to 19% less of the variance in predictees than do target models, on average.

4.2.3 Δ Correlation.

Another way to assess whether a model is faithful to data is to calculate the predicted correlations between variables in the model and see how well they compare with the actual correlations in the data. The rules for predicting correlations come from path analysis [16]: The *weight* of a path between two variables is the product of the weights on the links along the path, and the predicted correlation is the sum of the weights of all legal paths. For example, the predicted correlation between **2** and **7** in FBD-0 is the sum of the weights of three paths **2,6,7**; **6,8,7**; and **2,6,8,7**. If two variables point to a third and no path exists between them (e.g., **5,8** in FBD-0) then their predicted correlation is their actual correlation. We cannot calculate predicted correlations for the IC algorithm because it's unclear how to interpret undirected links. Δ Correlation is the mean absolute difference between predicted

and actual correlations for every pair of variables in a model. For FBD-0, Δ Correlation is very good, .036. The average score over all test models was not much higher, however, ranging from .047 to .065 as model size increases.

Because Δ Correlation summarizes differences between predicted and actual correlations, we can assess it for target models just as for FBD models. Surprisingly, Δ Correlation is actually higher for the target models than for FBD models, as shown in Table 3. If this seems impossible, recall that datasets are generated by randomly sampling variates in simultaneous equations that represent path models. Thus it is possible to get a dataset that doesn't correspond *very* well to the target model, and the opportunity arises for FBD to build a model that fits the data better than the target model does.

In fact we would argue that FBD-0 is a better model of its dataset than is TARGET-0, even though the data were generated from TARGET-0, and not only because of Δ Correlation: The link in FBD-0 from **4** to **1** is "wrong" in the sense that it doesn't appear in TARGET-0, but it is "right" in the sense that **4** accounts for much more variance in **1** than in **9**. It's worth keeping this example in mind as we consider other measures of FBD's performance, particularly the incidence of "correct" and "incorrect" links.

4.2.4 Total Links.

The mean numbers of links in 6-, 9-, and 12-variable *target* models were 10.8, 17.1, and 28.8, respectively. Some of these were correlation links; most were directed. The mean numbers of links in inferred models were, by model size, 9.18, 18.0, and 30.0, for FBD respectively; and 12.9, 26.1, and 43.2, respectively, for IC. All FBD links are directed; the figures for IC include all the links it found, including undirected links. Roughly seven links per IC model were directed, irrespective of the model size, so the number of undirected links in IC models rose from roughly 6 to 36 as the model size grew from 6 to 12 variables.

4.2.5 Correct Links, Incorrect Links.

A *correct link* is one that appears in both a target model and a corresponding inferred model. For example, the correct links for FBD-0 are $1 \rightarrow 9$, $8 \rightarrow 9$, $5 \rightarrow 9$, $7 \rightarrow 9$, $6 \rightarrow 8$, $2 \rightarrow 8$, $3 \rightarrow 6$. Note that FBD-0 should have a link between **4** and **9**. It failed to get one of the eight directed links in the target model, hence it's *Correct % Links* score is $7/8 = .875$. (Note that this statistic is not equal to the ratio of Correct Links over Total Links in the target model, because Total Links includes correlations. This statistic is correct directed links divided by total directed links.)

A link is *wrong* if it does not appear in the target model. In FBD-0, wrong links are $4 \rightarrow 1$, $4 \rightarrow 6$, $4 \rightarrow 7$, $4 \rightarrow 8$, $8 \rightarrow 7$, $5 \rightarrow 7$, $6 \rightarrow 7$, $2 \rightarrow 6$. For IC-0 the incorrect links are $5 \rightarrow 7$, $1 \rightarrow 8$, $4 \rightarrow 8$, $6 \rightarrow 3$. A link can be "wrong" in two ways: If the target model includes a link $X \rightarrow Y$ and an inferred model includes $Y \rightarrow X$, instead, then the latter is a *Wrong Reversed Link* (e.g., $6 \rightarrow 3$ in IC- 0). If an inferred model contains $Y \rightarrow X$ and the target model doesn't include a directed link from X to Y then the inferred link simply shouldn't be there, and it is called a *Wrong Not Reversed Link*.

An algorithm could score well on Correct Links by adding many links to a model, most of which are wrong. One measure of this tendency is the ratio of wrong links to correct links.

For example, for 6-variable models, FBD added .76 wrong links for every correct one; and for 9- and 12-variable models it added 1.18 and 1.66 wrong links, respectively, for every correct one. The IC algorithm has slightly higher numbers, 1.47, 1.5 and 2.17, respectively.

The story told by these numbers is familiar: a relatively conservative algorithm will make fewer mistakes and get fewer correct links than a more liberal one. The IC algorithm gets 3.3 links correct and 3.8 links wrong, on average; while for FBD the numbers are 8.7 and 10.3. Interestingly, the IC algorithm numbers do not depend on model size: it labels approximately seven links per model as directed, irrespective of model size. Consequently, the percentage of links it draws that are correct (*Correct%*) declines as models get bigger. The same is true of FBD, though the proportional effect is smaller.

4.2.6 Distance 1, 2 and 3.

When visually comparing two causal models, we might consider a missing link to be irrelevant if there is a directed path between the two variables in question. For example, FBD-0 is missing a link between **4** and **9**, but it includes a path from **4** to **1** to **9**. Thus, **4** does influence **9** in FBD-0, just not directly. The Distance 1 score is the percentage of directed links in a target model that are also in the corresponding inferred model. It is equal to *Correct%*. The Distance 2 score is the percentage of directed links in the target model for which there is a directed path of length 2 in the derived model. It is .125 for FBD-0 because one of eight direct links is represented by a path of length two, as described earlier. Similarly, the Distance 3 score tells us how many links in the target model are represented as paths of length 3 in the inferred model. When scoring the IC algorithm, we looked at directed links, only. Roughly 10% to 13% of the direct links in FBD models are represented by paths of length two or three; fewer for IC models, probably because they included fewer directed links.

4.2.7 Dependent Links, Colliders, Total.

These statistics measure whether an inferred model reflects the dependence relationships that hold between variables in the target model. We look at both direct and conditional dependencies in the target and inferred models. Pearl and Verma state, "[When all relevant variables are included in the model,] two causal models are equivalent iff their dags have the same links and the same set of uncoupled head-to-head nodes." [11] To measure this notion of equivalence, we compare each inferred model to the target model, and calculate number of correct links, the number of correct head-to-head (collider) nodes, and the total number of correct dependencies and colliders. (For IC, undirected links counted toward the dependence score, and directed links, only, counted toward the collider score.) Each score is then transformed to a percentage of the maximum number possible for the model; for a model with n variables, there are $n(n-1)/2$ such possibilities. FBD and IC both perform well and very similarly; they capture the dependency structure of the data with equal felicity.

5. Discussion of Experiment Results

We wanted to know two things: Does FBD build good *predictive* models, and how *accurate* are its models? Experiment 1 told us that when FBD must select k predictors, it chooses some but not all of the best k predictors. Consequently, its predictors account for 73% to 100% of the variance in the dependent variable that is accounted for by the best predictors. Experiment 1 was an "in vitro" study, however, because it didn't test FBD on complete models and it compelled FBD to select k predictors. Experiment 2 complements the results of Experiment 1: FBD's R^2 scores for the dependent variable were roughly 75% as high as those of the target models, and FBD models accounted for 15% less variance in all predictees, on average, than did the target models. In comparison, IC accounted for 35% less variance in all predictees, on average, not allowing undirected links, or 22% less, allowing undirected links.

As noted earlier, $\Delta Correlation$ scores represent the ability of a model to predict the empirical correlations among variables in a dataset. They were very good (i.e., low) for FBD. We think this probably is a consequence of using ω scores to select predictors. Recall that ω favors predictors for which β_{XY} is a big fraction of r_{XY}. It follows that predicted correlations, which are derived from β coefficients, are a big fraction of the actual correlations if the β coefficients are a big fraction of the correlations, as ω requires.

Recall, a set of predictors is not accepted unless it accounts for at least some fraction of the variance in the predictee (section 2.). For the previous results, this fraction was .1. When we raised it to .5, we got the data in Table 5. We did not expect the higher threshold to affect $DependentR^2$ but we were surprised by how little it affected $Nopenalty\Delta R^2$. With the higher threshold, FBD models accounted for 19% less variance in all predictees, on average, than did the target models.

The R^2 threshold had bigger effects on FBD's accuracy: With a low threshold (Table 4) FBD got 1.2 links wrong for every link it got right, on average; with the higher threshold (Table 5) it got .72 links wrong for every one correct. Not surprisingly, the higher threshold resulted in fewer links overall; the nice surprise was the increase in accuracy. As noted earlier, FBD got more correct links and also more wrong ones than the IC algorithm. This remains true even with the higher R^2 threshold.

We were particularly interested in the number of $WrongReversed$ links in FBD models. Suppose X_1 and X_2 are being considered as predictors of Y, and imagine X_2 predicts both Y and X_1 in the target model. Now, if X_2 is accepted as a predictor for Y but X_1 is not (because $\omega_{X_1Y} < T_\omega$) then when FBD looks for predictors for X_2, it will probably select X_1. (If X_2 causes X_1, then X_1 will often be a good predictor of X_2.) This phenomenon will show up as a $WrongReversed$ link. Forty-three percent of wrong links were reversed in FBD 6-variable models (Table 4). This proportion dropped to 26% and 27% for 9- and 12-variable models, respectively. Evidently, this phenomenon is a weakness of the FBD algorithm, but we have some evidence that it doesn't get worse as models get bigger and, in any case, it accounts for less than half of the links FBD gets wrong. These results are affected little by increasing the R^2 threshold (Table 5), which isn't surprising, considering that high thresholds discard entire sets of predictors, whereas this phenomenon is observed when one predictor from a set is discarded incorrectly.

A different view of accuracy is given by *Dependencies* and *Colliders* scores, which rep-

resent how well a model accounts for the dependencies and conditional dependencies in a dataset. Overall, FBD performed as well or better than IC on this measure, and the figures were affected little by the R^2 threshold.

Linear regression is not popular, currently, with the causal induction community. One reason is that many relationships are not linear, but this concern can often be addressed by transforming one's data beforehand [5]. A more subtle issue, raised by Glymour, Spirtes and Scheines, is that beta coefficients are unstable, especially when unmeasured or *latent* variables influence them. Selecting variables by their ω scores lessens this problem. We have obtained good results with models Glymour et al. [14, page 240] show are difficult for ordinary regression; see [1].[3] Nevertheless, most causal induction algorithms attempt to build models that are consistent with conditional independence relationships, and they discard quantitative covariance information once these relationships are inferred.

Yet FBD, based on linear regression, performed as well as IC, based on conditional independence, when we measured its ability to account for dependencies in the data. One reason is that ω scores provide much the same information as conditional independence tests: Because βs are partial, low ω is akin to finding high correlation and low partial correlation, which is basically the conditional independence test in IC. In addition to its strong dependency scores, though, FBD also built strongly predictive models. We are not claiming FBD is "better" than IC because they solve different problems. Surely, though, these results suggest regression deserves a closer look.

[3] FBD chose the correct predictors 88% of the time. Variables from the set of correct predictors were always included; the remaining 12% is due to not choosing the entire set of correct predictors. When FBD could have incorrectly chosen variables whose relationships with the dependent variable were due to latent variables or correlations with predictors, they were rejected 82% of the time. 39% of those rejections were due to ω pruning.

k	$Vars.$	$\mid p-batch(k) \cap p-best(k) \mid$	$\mid p-iter.(k) \cap p-best(k) \mid$
5	12	3.15 (1.16)	3.375 (.86)
5	9	3.65 (.49)	3.7 (.52)
5	6	5.0 (0)	5.0 (0)
4	12	2.3 (.57)	2.3 (.74)
4	9	3.05 (.36)	3.08 (.53)
4	6	3.33 (.23)	3.33 (.23)
3	12	1.48 (.61)	1.68 (.58)
3	9	2.18 (.46)	2.18 (.51)
3	6	2.28 (.36)	2.33 (.33)

Table 1: Means and (Standard Deviations) of the size of the intersections

k	$Vars.$	$R^2_{p-batch(k)}/R^2_{p-best(k)}$	$R^2_{p-iter.(k)}/R^2_{p-best(k)}$
5	12	.845 (.03)	.86 (.023)
5	9	.94 (.006)	.94 (.005)
5	6	1.0 (0)	1.0 (0)
4	12	.80 (.04)	.82 (.03)
4	9	.94 (.008)	.94 (.005)
4	6	.91 (.01)	.91 (.01)
3	12	.73 (.06)	.80 (.04)
3	9	.91 (.02)	.92 (.01)
3	6	.88 (.03)	.89 (.03)

Table 2: Means and (Standard Deviations) of the R^2 ratios

$Measure$	$6vars$	$9vars$	$12vars$
$DependentR^2$	.768 (.163)	.833 (.109)	.832 (.168)
$\Delta Corr.$	.064 (.054)	.063 (.054)	.098 (.089)
$TotalLinks$	10.8	17.1	28.8

Table 3: Target Model Statistics

	FBD			IC		
Measure	*6vars*	*9vars*	*12vars*	*6vars*	*9vars*	*12vars*
Dependent R^2	.74 (.189)	.795 (.145)	.745 (.276)	.326 (.389)	.451 (.338)	.303 (.363)
Dep. R^2 *w/corr.*				.603 (.31)	.811 (.12)	.739 (.216)
Nopenalty ΔR^2	.13 (.089)	.137 (.081)	.193 (.126)	.347 (.128)	.345 (.137)	.348 (.095)
NoPen. R^2 *w/corr.*				.199 (.104)	.124 (.063)	.172 (.063)
Δ*Corr.*	.047 (.02)	.054 (.03)	.065 (.03)			
TotalLinks	9.18	18.0	30.0	12.9	26.1	43.2
CorrectLinks	5.6 (1.60)	8.85 (2.96)	11.6 (4.11)	2.65 (2.06)	3.4 (2.19)	3.75 (2.71)
Correct%	.666 (.192)	.681 (.164)	.543 (.221)	.293 (.18)	.272 (.19)	.165 (.11)
WrongLinks	3.6 (1.19)	9.1 (2.32)	18.3 (5.86)	2.75 (1.41)	3.85 (2.37)	4.9 (2.49)
WrongReversed	1.55 (1.28)	2.4 (1.63)	5.0 (3.69)	1.55 (1.05)	1.35 (1.27)	2.0 (1.30)
WrongNotRev.	2.05 (1.23)	6.7 (2.51)	13.3 (5.74)	1.2 (1.06)	2.5 (2.09)	2.9 (1.8)
Wrong/Correct.	.76 (.55)	1.18 (.61)	1.66 (.83)	1.48 (.944)	1.50 (1.11)	2.17 (2.10)
Distance1	.666 (.193)	.682 (.166)	.543 (.221)	.293 (.180)	.272 (.194)	.165 (.108)
Distance2	.092 (.099)	.084 (.071)	.107 (.071)	.016 (.041)	.013 (.028)	.031 (.041)
Distance3	.013 (.039)	.015 (.036)	.023 (.030)	0 (0)	.005 (.020)	.004 (.012)
Dependencies	.667 (.155)	.722 (.1)	.691 (.09)	.630 (.109)	.739 (.095)	.707 (.077)
Colliders	.757 (.13)	.721 (.08)	.689 (.071)	.707 (.134)	.756 (.071)	.741 (.095)
Total	.783 (.17)	.693 (.109)	.626 (.105)	.657 (.161)	.642 (.095)	.595 (.1)

Table 4: Means and (Standard Deviations) of the results from Experiment 2

Measure	*6vars*	*9vars*	*12vars*
Dependent R^2	.723 (.235)	.795 (.146)	.725 (.313)
Nopenalty ΔR^2	.175 (.097)	.167 (.087)	.235 (.123)
Δ*Corr.*	.046 (.037)	.057 (.049)	.065 (.036)
TotalLinks	6.35	12.95	23.1
CorrectLinks	4.55 (2.28)	8.2 (3.24)	10.45 (4.89)
Correct%	.53 (.25)	.62 (.16)	.46 (.22)
WrongLinks	1.8 (1.64)	4.75 (2.81)	12.65 (6.38)
WrongReversed	.95 (1.23)	1.35 (1.66)	4.35 (3.69)
WrongNotRev.	.85 (1.04)	3.4 (1.76)	8.3 (3.71)
Wrong/Correct.	.4 (.48)	..63 (.42)	1.2 (.76)
Distance1	.535 (.247)	.619 (.165)	.462 (.219)
Distance2	.041 (.08)	.073 (.071)	.088 (.081)
Distance3	.006 (.025)	.006 (.018)	.015 (.026)
Dependencies	.62 (.17)	.75 (.10)	.70 (.09)
Colliders	.76 (.13)	.75 (.10)	.70 (.09)
Total	.69 (.27)	.71 (.12)	.64 (.09)

Table 5: Results from Expt 2 $R^2Threshold = .5$

Acknowledgments

This research is supported by ARPA under contract F30602-93-C-0100, and by a NASA GSRP Training Grant, #NGT-70358. We thank Dr. Mike Sutherland and Prof. Glenn Shafer for helpful discussions. Prof. Judea Pearl kindly let us use the IC algorithm. Prof. Clark Glymour provided algorithms from the TETRAD family, as well as helpful advice. We also thank Adam Carlson, David Hart and David Westbrook for discussions, advice and support for CLASP/CLIP.

Availability of Code

FBD is available as part of the CLASP/CLIP package. If interested, please contact David Hart (hart@cs.umass.edu).

REFERENCES

[1] Lisa Ballesteros. Regression–based causal induction with latent variable models. Submitted to AAAI, 1994.

[2] P. Bentler. *Theory and Implementation of EQS: A Strutural Equations Program.* BMDP Statistical Software, Inc., Los Angeles, 1985.

[3] Paul R. Cohen, Lisa Ballesteros, Dawn Gregory, and Robert St. Amant. Experiments with a regression–based causal induction algorithm. Unpublished technical report, 1994.

[4] G. Cooper and E. Herskovits. A bayesian method for constructing bayesian belief networks from databases. In Bruce D'Ambrosio, Philippe Smets, and Piero Bonissone, editors, *Uncertainty in Artificial Intelligence*, pages 86–94. Morgan Kaufmann, 1991.

[5] David Hoaglin, Frederick Mosteller, and John Tukey. *Understanding robust and exploratory data analysis.* John Wiley and Sons, Inc., 1983.

[6] Paul Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

[7] Y. Iwasaki and H.A. Simon. Causality in device behavior. *Artificial Intelligence*, 29:3–32, 1986.

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

[9] C.C. Li. *Path analysis-a primer.* Boxwood Press, 1975.

[10] Judea Pearl and T.S. Verma. A statistical semantics for causation. *Statistics and Computing*, 2:91–95, 1991.

[11] Judea Pearl and T.S. Verma. A theory of inferred causation. In J. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference.*, pages 441–452. Morgan Kaufman, 1991.

[12] H. Simon. Spurious correlations: A causal interpretation. *Journal of the American Statistical Association*, 49:469–492, 1954.

[13] Robert R. Sokal and F. James Rohlf. *Biometry: the principles and practice of statistics in biological research.* W.H. Freeman and Co., New York, second edition, 1981.

[14] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer-Verlag, 1993.

[15] P.C. Suppes. *A Probabilistic Theory of Causality.* North Holland, Amsterdam, 1970.

[16] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

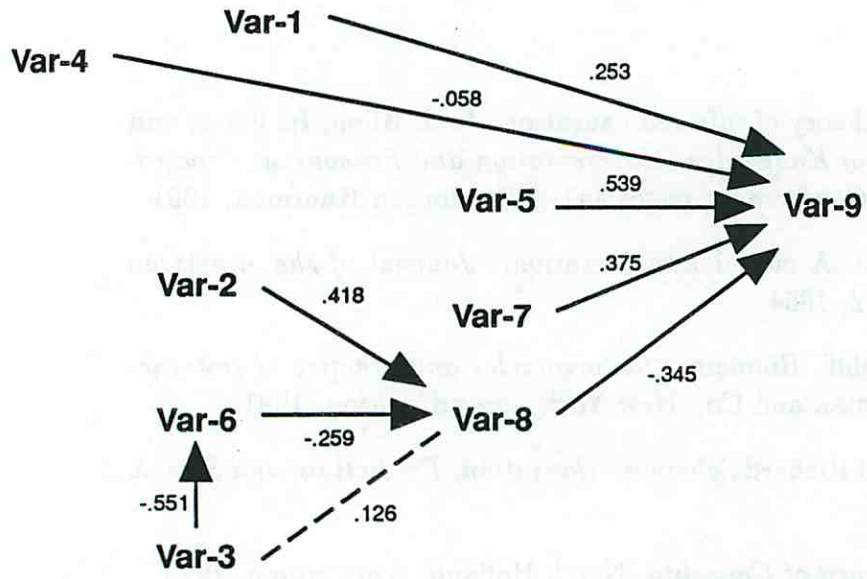

Figure 1.a: Target-0

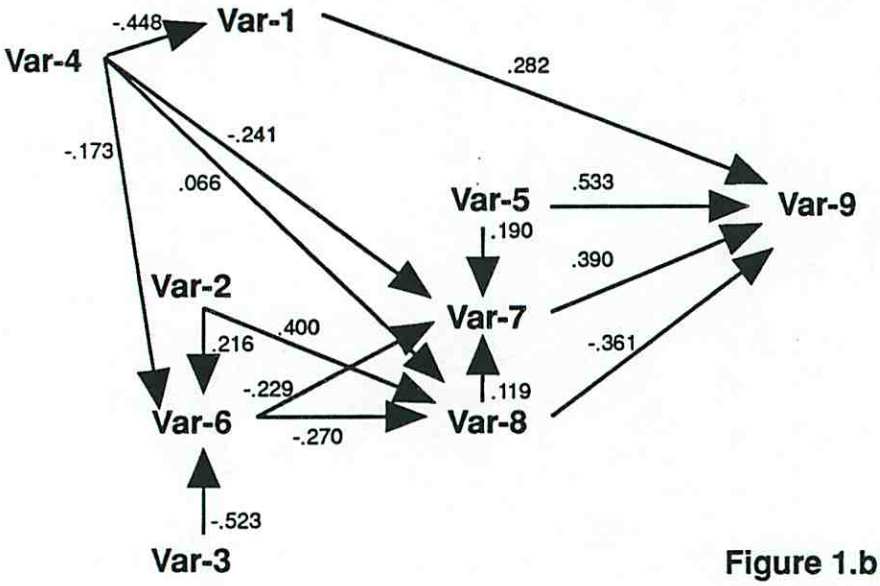

Figure 1.b: FBD-0

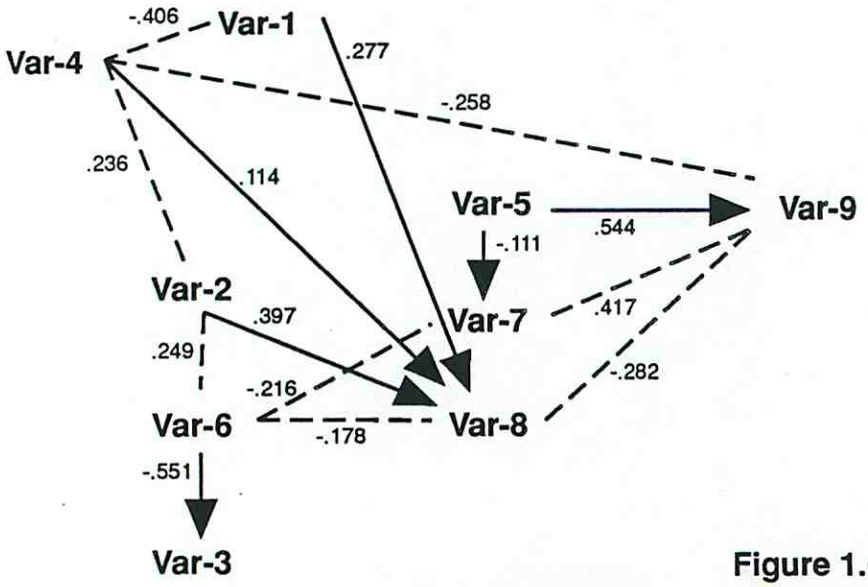

Figure 1.c: IC-0