Measurement Error and Causal Discovery

Richard Scheines & Joseph Ramsey
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15217, USA

1 Introduction

Algorithms for causal discovery emerged in the early 1990s and have since proliferated [4, 10]. After directed acyclic graphical representations of causal structures (causal graphs) were connected to conditional independence relations (the Causal Markov Condition[1] and d-separation[2]), graphical characterizations of Markov equivalence classes of causal graphs (patterns) soon followed, along with pointwise consistent algorithms to search for patterns. Researchers in Philosophy, Statistics, and Computer Science have produced constraint-based algorithms, score-based algorithms, information-theoretic algorithms, algorithms for linear models with non-Gaussian errors, algorithms for systems that involve causal feedback, algorithms for equivalence classes that contain unmeasured common causes, algorithms for time-series, algorithms for handling both experimental and non-experimental data, algorithms for dealing with datasets that overlap on a proper subset of their variables, and algorithms for discovering the measurement model structure for psychometric models involving dozens of "indicators". In many cases we have proofs of the asymptotic reliability of these algorithms, and in almost all cases we have simulation studies that give us some sense of the finite-sample accuracy of these algorithms. The FGES algorithm (Fast Greedy Equivalence Search, [6]), which we feature here, is highly accurate in a wide variety of circumstances and is computationally tractable on a million variables for sparse graphs. Many algorithms have been applied to serious scientific problems like distinguishing between Autistic and neurotypical subjects from fMRI data [2], and interest in the field seems to be exploding.

Amazingly, work to assess the finite-sample reliability of causal discovery algorithms has proceeded under the assumption that the variables given are measured without error. In almost any empirical context, however, some of the variance of any measured variable comes from instrument noise or recording errors or worse. Historically, the development of statistics as a discipline was partly spurred by the need to handle measurement error in astronomy. In many cases measurement error can be substantial. For example, in epidemiology, cumulative exposure to environmental pollutants or toxic chemicals is often measured by proxies only loosely correlated with exposure, like "distance from an industrial pollutant emitter" – or an air quality monitor within a few miles of the subject.

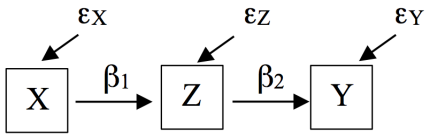
In its simplest form, measurement error is random noise: $X_{\text{measure}} = X + \epsilon$, where ϵ is as an aggregate representing many small but independent sources of error and thus by the central limit theorem at least approximately Gaussian. In other cases measurement error is systematic, for example it is well known that people under-report socially undesirable activities like cheating, and since the more they engage in the activity the more they under-report, this type of error is not random. In this paper we will concern ourselves only with random noise error. Here we explore the impact of random noise measurement error on the overall accuracy of causal discovery algorithms.

2 Parameterizing Measurement Error

We consider linear structural equation models (SEMs) in which each variable V is a linear combination of its direct causes and an "error" term ϵ_V that represents an aggregate of all other causes of V. In Figure 1, we show a simple model involving a causal chain from X to Z to Y. Each variable has a structural equation, and the model can be parameterized by assigning real values to β_1 and β_2, and a joint normal distribution to $\{\epsilon_X, \epsilon_Z, \epsilon_Y\} \sim N(0, \Sigma^2)$, with Σ^2 diagonal to reflect the independence among the "error terms" ϵ_X, ϵ_Z and ϵ_Y.

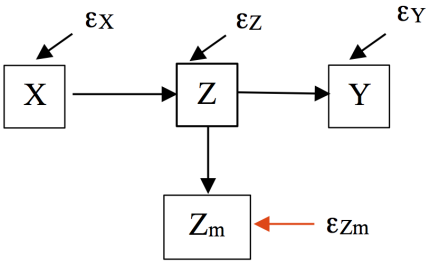
For any values of its free parameters, the model in Fig. 1 entails the vanishing partial correlation $\rho_{XY.Z} = 0$, which in the Gaussian case is also the conditional independence: $X \perp\!\!\!\perp Y \mid Z$. In Fig. 2 we show the same model, but with Z "measured" by Z_m, with "measurement error" ϵ_{Z_m}.

[1] Spirtes et al. [10]
[2] Pearl [4]



Equations:
$$X := \epsilon_X$$
$$Z := \beta_1 X + \epsilon_Z$$
$$Y := \beta_2 Z + \epsilon_Y$$

Figure 1: Causal Model for X, Z, Y.



Equations:
$$X := \epsilon_X$$
$$Z := \beta_1 X + \epsilon_Z$$
$$Y := \beta_2 Z + \epsilon_Y$$
$$Z_m := Z + \epsilon_{Z_m}$$

Figure 2: Measurement Error on Z.

The variance of Z_m is just the sum of the variances of Z and the measurement error ϵ_{Z_m}, i.e., $var(Z_m) = var(Z) + var(\epsilon_{Z_m})$, so we can parameterize the "amount" of measurement error by the fraction of $var(Z_m)$ that comes from $var(\epsilon_{Z_m})$. If $var(\epsilon_{Z_m}) = 0.5var(Z_m)$, then half of the variation in the measure is from measurement noise. For simplicity, and without loss of generality, consider the re-parameterization of Fig. 2 in which all variables $\{X, Z, Z_m, Y\}$ are standardized to have mean 0 and variance 1. In that case $Z_m = \lambda Z + \epsilon_{Z_m}$, where $\rho_{Z,Z_m} = \lambda^2$ and $var(\epsilon_{Z_m}) = 1 - \lambda^2$.[3]

3 The Problem for Causal Discovery

Measurement error presents at least two sorts of problems for causal discovery. First, for two variables X and Y that are measured with error by X_m and Y_m, if X and Y have correlation ρ_{XY}, then this correlation will attenuate in the measures X_m and Y_m. That is, $|\rho_{XY}| > |\rho_{X_m Y_m}|$.

Thus it might be the case that a sample correlation ρ_{XY} would lead us to believe that X and Y are correlated in the population, i.e., $|\rho_{XY}| > 0$, but the sample correlation ρ_{X_m,Y_m} would mislead us into rejecting independence and accepting $\rho_{X_m Y_m} = 0$. Thus an algorithm that would make X and Y adjacent because a statistical inference concludes that $|\rho_{XY}| > 0$, the same procedure might conclude that X_m and Y_m are not adjacent because a statistical inference concludes that $\rho_{X_m Y_m} = 0$. Worse, this decision will in many cases affect other decisions that will in turn affect the output of the procedure.

Second, if a variable Z "separates" two other variables X and Y, that is X and Y are d-separated by Z and thus $X \perp\!\!\!\perp Y \mid Z$, then a measure of Z with error Z_m does not separate X and Y, i.e., $X \not\perp\!\!\!\perp Y \mid Z_m$. In Fig. 2, for example, the model implies that $X \perp\!\!\!\perp Y \mid Z$, but it does not imply that $X \perp\!\!\!\perp Y \mid Z_m$. If the measurement error is small, we still might be able to statistically infer that $X \perp\!\!\!\perp Y \mid Z_m$, but in general measurement error in a separator fails to preserve conditional independence. Again, judgments regarding $X \perp\!\!\!\perp \mid Z_m$, will affect decisions involving relationships between X, Y, Z_m , but it will also have non-local consequences involving other variables in the graph.

For example, consider a case in which we simulated data from a model with six variables in which only two of the six were measured with error. In Fig. 3, the generating model on the left is a standardized SEM (all variables mean 0 and variance 1), with $L1$ and $L2$ measured with 1% error as $L1m$ and $L2m$. Running FGES on a sample of size $1,000$ drawn from the measured variables $\{L1m, L2m, X1, X2, X3, X4\}$, we obtained the pattern output on the right, which is optimal even if we had the population distribution for $\{L1, L2, X1, X2, X3, X4\}$. Measuring $L1$ and $L2$ with 20% error produced a pattern nothing like the true one (Fig. 4), and the errors in the output are not restricted to errors involving relations involving $L1m$ and $L2m$ as potential separators.

For example, FGES found, in error, that $X1$ and $X4$ are adjacent because $L2m$ did not separate them where $L2$ would have. Similarly, $X1$ and $X2$ were made adjacent because $L1m$ failed to separate them where $L1$ would have. $X2$ and $X4$ were found to be separated, but not by $X1$, thus the algorithm oriented the false adjacencies $X2$ — $X1$ and $X4$ — $X1$ as $X2 \rightarrow X1$ and $X4 \rightarrow X1$, which in turn caused it to orient the $X1$ — $L1m$ adjacency incorrectly as $X1 \rightarrow L1$. Thus errors made as a result of measurement error are not contained to local decisions. Because of this non-locality, judging the overall problem for causal discovery algorithms posed by measurement error is almost impossible to handle purely theoretically.

4 Causal Discovery Accuracy

Causal researchers in several fields recognize that measurement error is a serious problem, and strategies have

[3] $var(Z_m) = \lambda^2 var(Z) + var(\epsilon_{Z_m}) = \lambda^2 + var(\epsilon_{Z_m}) = 1$, so $var(\epsilon_{Z_m}) = 1 - \lambda^2$.

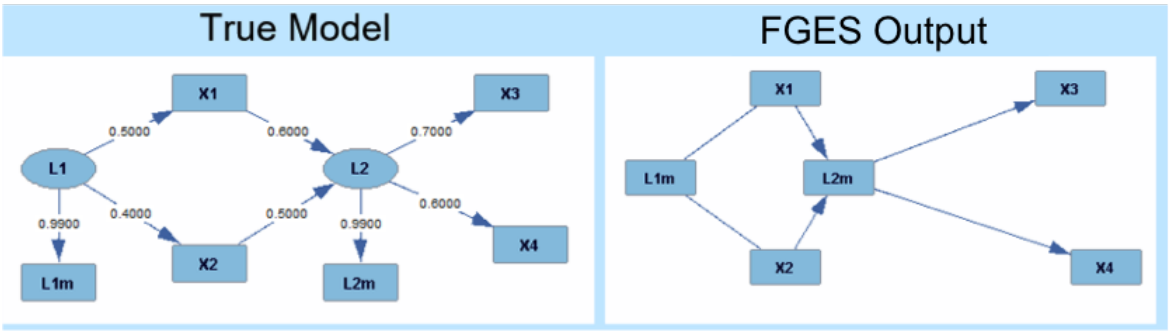


Figure 3: FGES Output with 1% Measurement Error.

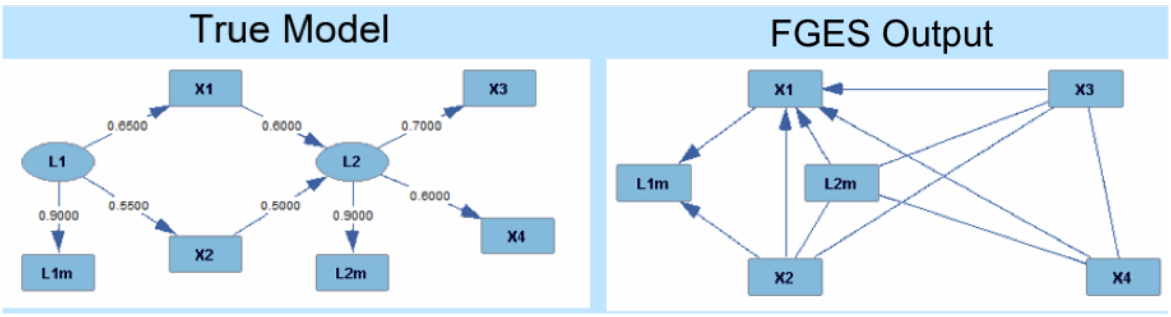


Figure 4: FGES Output with 20% Measurement Error.

emerged for dealing with the problem theoretically. In multivariate regression, it is known that a) measurement error in the response variable Y inflates the standard errors of the coefficient estimates relating independent variables to Y but does not change their expectation, b) measurement error in an independent variable X attenuates the coefficient estimate for X toward 0, and c) measurement error in a "covariate" Z produces partial omitted variable bias in any estimate on a variable X for which Z is also included in the regression. In cases b) and c), there is a literature, primarily in economics, known as "errors-in-variables", for handling measurement error in the independent variables and covariates [1]. If the precise amount of measurement error is known, then parameter estimates can be adjusted accordingly. If the amount is unknown, then one can still use sensitivity analysis or a Bayesian approach [7]. In general, in cases in which researchers are confident in the specification of a model, the effect of measurement error on parameter estimation has been well studied.[4]

In causal discovery algorithms, however, the input is typically assumed to be an i.i.d. sample for a set of measured variables $\mathbf{V}$ drawn from a population with variables $\mathbf{V} \cup \mathbf{L}$ that satisfies some general assumptions (e.g., Causal Markov and/or Faithfulness Axiom) and/or some parametric assumption (e.g., linearity). The output is often an equivalence class of causal structures – any representative of which might be an identified statistical model whose parameters can be estimated from these same data. In discussing the reliability of the causal discovery algorithm, we care about how the equivalence class output by the algorithm compares to the equivalence class of which the true causal graph is a member.

For example, if the causal graph in the upper left of Fig. 5 is the true model, then the pattern on the upper right represents the Markov equivalence class of which the true graph is a member. It represents all that can be discovered about the causal structure under the assumption that the evidence is non-experimental, confined to independence relations, and the connection between the graph and data is the Causal Markov and Faithfulness assumptions. On the bottom of Fig. 5 we show the patterns output by the FGES algorithm on a sample of 50 (bottom left), and on a sample of 1,000 (bottom right). The pattern on the bottom right matches the pattern for the true graph, so this output is maximally accurate, even though it did not "orient" the edge between $X2$ and $X4$.

Patterns output by search procedures can be scored on accuracy with respect to adjacencies and accuracy with respect to arrowhead orientation. The true pattern has three adjacencies: $\{X1 - X3,\ X2 - X3,\ X2 - X4\}$, and two arrowhead orientations: $\{X1 \rightarrow X3, X2 \rightarrow X3\}$. The pattern on the bottom left contains two of the three adjacencies, but it missed the adjacency between $X2$ and $X3$. The adjacency precision (AP) reflects the proportion, among those guessed to be adjacencies, of correct adja-

[4]For example, see Pearl [5], and Kuroki and Pearl [3].

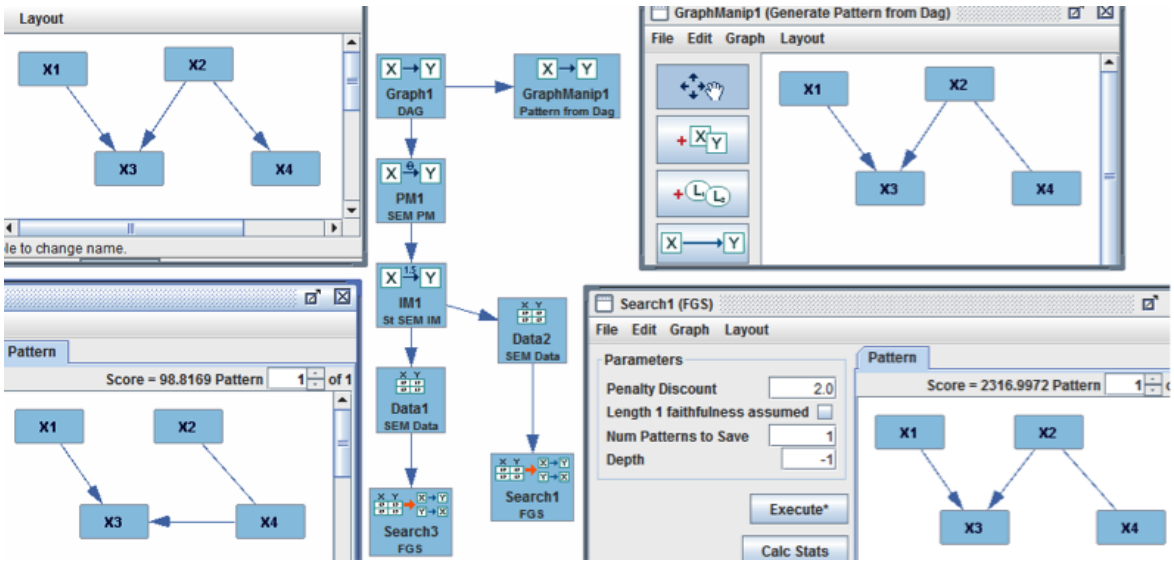


Figure 5: Graphs, Patterns, and Search Accuracy

cencies. In this case the search algorithm guessed three adjacencies, two of which were correct, so the adjacency precision is $2/3 = .67$. The adjacency recall (AR) is the proportion of true adjacencies found by the algorithm. In this case the search output two of three, so $AR = .67$.

We can also compute the "arrowhead precision" (AHP) and "arrowhead recall" (AHR). In the bottom left of Fig.3 the algorithm output two arrowheads, but we only count the one which also corresponds to a true adjacency, so its AHP is $1/1 = 1.0$. As there were two arrowheads in the true pattern that it could have found, one of which it missed, its $AHP = 1/2 = .5$.

5 The Simulation Study

As predicting the global effects of measurement error on search accuracy is hopelessly complicated by the non-locality of the problem, we instead performed a somewhat systematic simulation study. We generated causal graphs randomly, parameterized them randomly, drew pseudo-random samples of varying size, added varying degrees of measurement error, and then ran the FGES causal discovery algorithm and computed the four measures of accuracy we discussed above: AP, AR, AHP, and AHR.

We generated random causal graphs, each with 20 variables. In half of our simulations the "average degree" of the graph (the average number of adjacencies for a variable) was 2 and in half it was 6. In a graph with 20 variables with an average degree of 2, the graph is fairly sparse. With an average degree of 6 the graph is fairly dense. We randomly chose edge coefficients, and for each SEM we generated 10 samples of size 100, 500, 1,000, and 5,000. For each sample we standardized the variables to have mean 0 and variance 1, and we then applied varying degrees of measurement error to all of the measured variables. For each variable X, we replaced X with $X_m = X + \epsilon_x$, where $\epsilon_x \sim N(0, \sigma^2)$, for values of $\sigma^2 \in \{0, .01, .1, .2, .3, .4, .5, 1, 2, 3, 4, 5\}$. Since the variance of all the original variables is 1.0, when $\sigma^2 = .1$ approximately 10% of the variance of X_m was from measurement error. When $\sigma^2 = 1$, 50% of the variance of X_m was from measurement error, and when $\sigma^2 = 5$, over 83% of the variance of X_m was measurement error.

6 Results

First consider the effect of sample size on FGES accuracy when there is no measurement error at all. In Fig. 6 we show the performance for sparse graphs (average degree = 2) and for fairly dense graphs (average degree = 6). For sparse graphs, accuracy is high even for samples of only 100, and by $N = 1,000$ is almost maximal.

For denser graphs, which are much harder to discover, performance is still not ideal even at sample size 5,000. Fig. 7 shows the effects of measurement error on accuracy for both sparse and dense graphs at sample size 100, 500, and 5,000.

For sparse graphs, at sample size 100 the of accuracy of FGES decays severely for measurement errors of less than 17% (ME = .2), especially for orientation accuracy. The results are quite poor at 50% (ME = 1) and reaches near 0 when two thirds of the variance of each variable is noise. Surprisingly, the decay in performance of FGES from measurement error is much slower at larger sample sizes. For $N = 500$, the algorithm is still reasonably accurate at 50% measurement error, and adjacency precision remains rea-

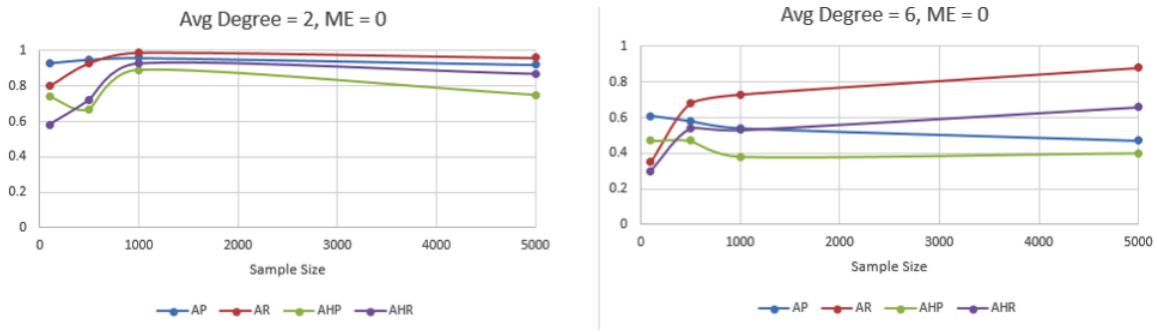


Figure 6: Baseline: Zero Measurement Error

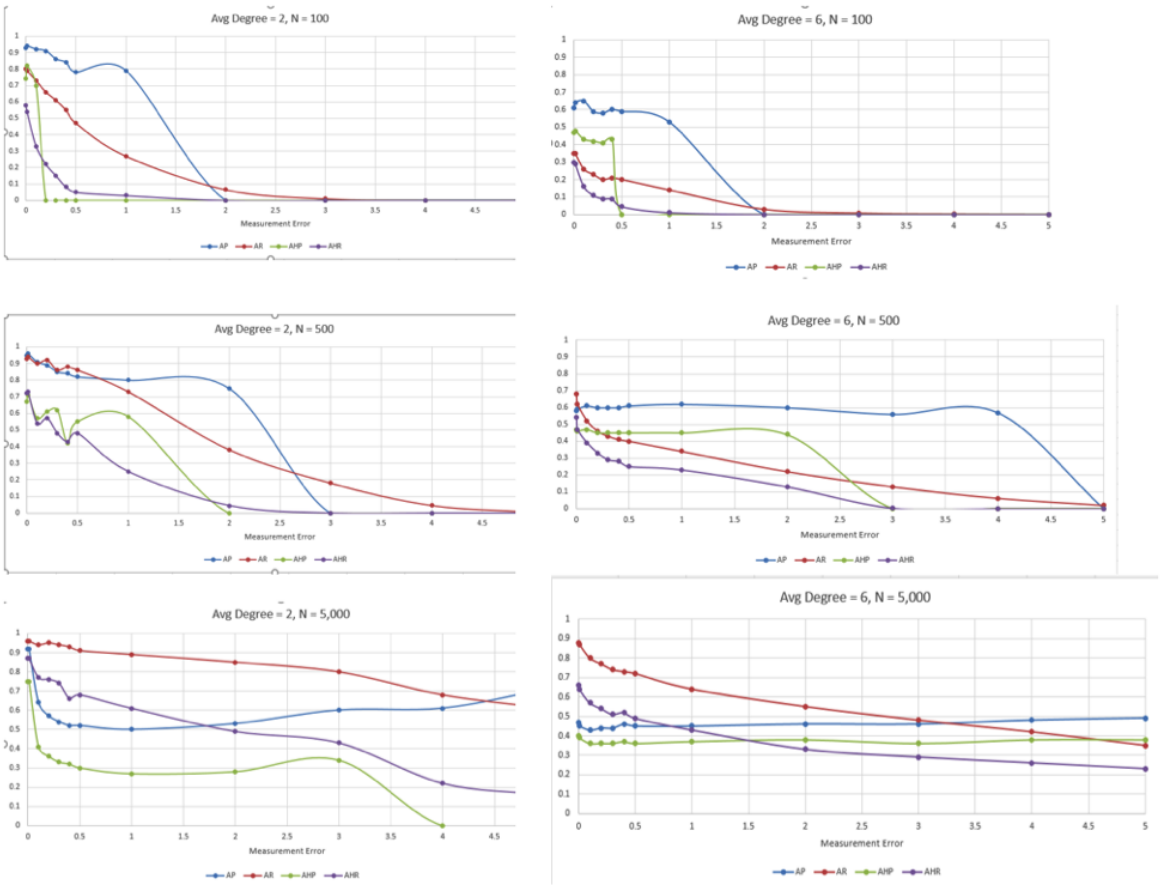


Figure 7: Accuracy and Measurement Error

sonably high even when $2/3$ of the variance of all variables is just noise. If FGES outputs an adjacency at sample size 500, it is still over 70% likely to actually be an adjacency in the true pattern, even with 67% measurement error. On the other hand, the same can't be said about non-adjacencies (AR) in 67% measurement error case. If FGES outputs a pattern in which X and Y are not adjacent, then there is a lower than 40% chance that X and Y are not adjacent in the true pattern – even in patterns with average degree of 2.

Does sample size help FGES's performance in the presence of measurement error? Roughly yes. For dense graphs the performance of FGES improves somewhat with sample size at almost all levels of measurement error. Notably, even for cases in which $5/6$ of the variance in the measured variables was due to measurement error (ME = 5), most measures of FGES accuracy improved dramatically when sample size was increased from $N = 500$ to $N = 5,000$.

7 Discussion, Conclusions and Future Research

We explored the effect of the simplest form of Gaussian measurement error on FGES, a simple causal discovery algorithm, in a simple causal context: no latent confounders, all relations are linear, and the variables are distributed normally. In this context, even small amounts of measurement error seriously diminish the algorithm's accuracy on small samples, especially its accuracy with respect to orientation. Surprisingly, at large sample sizes ($N = 5,000$) FGES is still somewhat accurate even when over 80% of the variance in each variable is random noise.

One way to conceive of the situation we describe is as latent variable model discovery, and use discovery algorithms built to search for PAGs (partial ancestral graphs that represent models that include latent confounding) instead of patterns. As the variables being measured with error are in some sense latent variables, this is technically appropriate. The kind of latent confounding for which PAGs are meant, however, is not the kind typically produced by measurement error on individual variables. If the true graph, for example, is $X \leftarrow Y \rightarrow Z$, and we measure X, Y, and Z with error, then we still want the output of a PAG search to be: X_m o–o Y_m o–o Z_m, which we would interpret to mean that any connection between X_m and Z_m goes through Y_m (and by proxie Y). If measurement error caused a problem, however, then the output would likely be a PAG with every pair of variables connected by the "o–o" adjacency, a result that is technically accurate but which carries almost no information.

There are two strategies that might work and that we will try on this problem in future research. First, prior knowledge about the degree of measurement error on each variable might be used in a Bayesian approach in which we compute a posterior over each independence decision. This would involve serious computational challenges, and we do not know many variables would be feasible with this approach. It is reasonable in principle, and a variant of it was tried on the effect of Lead on IQ among children [8].

Second, if each variable studied is measured with at least two "measurement" variables with independent measurement error, then the independence relations among the "latent" variables in a linear system can be directly tested by estimating a structural equation model and performing a hypothesis test on a particular parameter, which is 0 in the population if and only if the independence holds in the population.[5] This strategy depends on the measurement error in each of the measures being independent, an assumption that is often false. Techniques to find measurement variables that do have independent error have been developed, however [9], and these techniques can be used to help identify measures with independent error in situations where measurement is a serious worry.

A different measurement problem that is likely to produce results similar to Gaussian measurement noise is discretization. We know that, if, for continuous variables X, Y, and Z, $X \perp\!\!\!\perp Y \mid Z$, then for discrete projections of X, Y, and Z: X_d, Y_d, and Z_d, in almost every case the independence fails. That is, it is often the case that $X \perp\!\!\!\perp Y \mid Z$ but $X_d \not\perp\!\!\!\perp Y_d \mid Z_d$. This means that using discrete measures of variables that are more plausibly continuous likely leads to the same problems for causal discovery as measuring continuous variables with error. As far as we are aware, no one has any clear idea of the severity of the problem, even though using discrete measures of plausibly continuous variables is commonplace.

Acknowledgments

This research was supported by NIH grant #U54HG008540 (the Center for Causal Discovery), and by NSF grant #1317428.

References

[1] Wayne A. Fuller. *Measurement Error Models*. John Wiley & Sons, 1987.

[2] C. Hanson, S. J. Hanson, J. Ramsey, and C. Glymour. Atypical effective connectivity of social brain networks in individuals with autism. *Brain Connectivity*, 3(6):578–589, 2013.

[3] M. Kuroki and J. Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

[4] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[5] J. Pearl. Measurement bias in causal inference. *Proceedings of 26th Conference in Uncertainty in Artificial Intelligence*, pages 425–432, 2010.

[6] J. Ramsey. Scaling up greedy causal search for continu-

[5] See chapter 12, section 12.5 in Spirtes et al. [11].

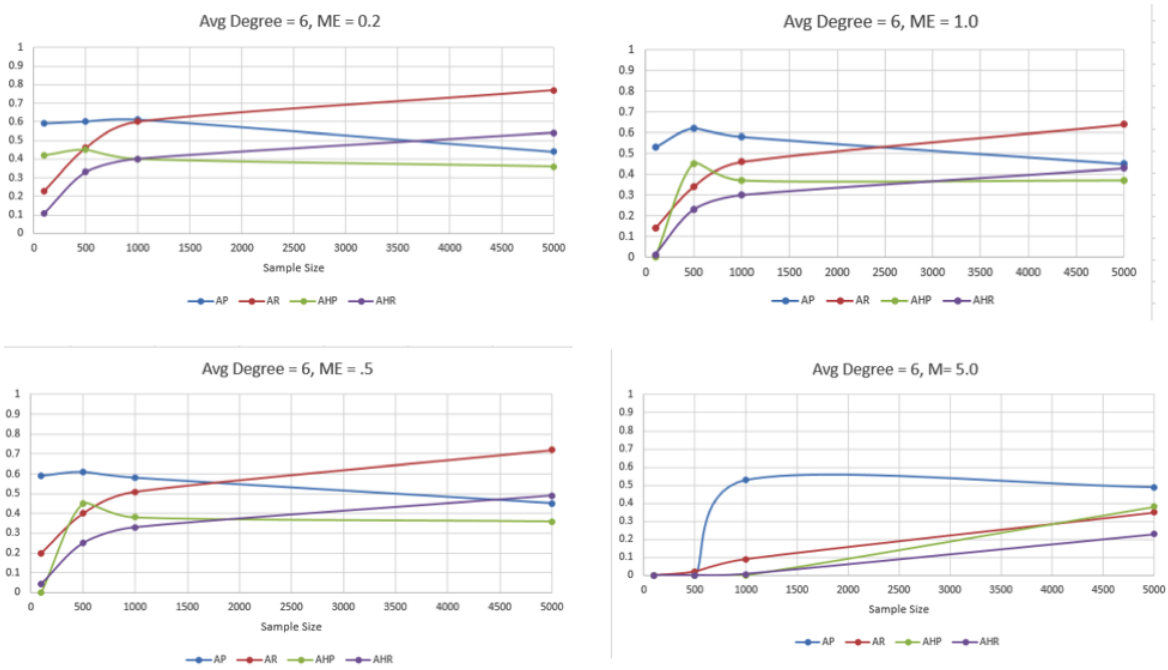


Figure 8: Sample Size vs. Accuracy in the presence of measurement error.

ous variables. Technical report, Center for Causal Discover, 2015. arXiv:1507.07749.

[7] P. Rosenbaum. *Observational Studies*. Springer-Verlag, 1995.

[8] R. Scheines. Estimating latent causal influences: Tetrad III variable selection and Bayesian parameter estimation: the effect of lead on IQ. In *Handbook of Data Mining*. Oxford University Press, 2000.

[9] R. Silva, C. Glymour, R. Scheines, and P. Spirtes. Learning the structure of latent linear structure models. *Journal of Machine Learning Research*, 7:191–246, 2006.

[10] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search. Lecture Notes in Statistics, 81*. Springer-Verlag, 1993.

[11] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, 2000.