A Study of Causal Discovery With Weak Links and Small Samples

Honghua Dai†, Kevin Korb†, Chris Wallace*, Xindong Wu‡
†Dept of Computer Science ‡Dept of Software Development
Monash University, Clayton, Victoria 3168, AUSTRALIA
dai@bruce.cs.monash.edu.au

Abstract

Weak causal relationships and small sample size pose two significant difficulties to the automatic discovery of causal models from observational data. This paper examines the influence of weak causal links and varying sample sizes on the discovery of causal models. The experimental results illustrate the effect of larger sample sizes for discovering causal models reliably and the relevance of the strength of causal links and the complexity of the original causal model. We present indicative evidence of the superior robustness of MML (Minimum Message Length) methods to standard significance tests in the recovery of causal links. The comparative results show that the MML-CI (the MML Causal Inducer) causal discovery system finds better models than TETRAD II given small samples from linear causal models. The experimental results also reveal that MML-CI finds weak links with smaller sample sizes than can TETRAD II.

1 Introduction

Our research on automating causal discovery aims at developing methods of reliably recovering the structure (and parameters) of causal models from sample data. Given such a method, several factors will affect the correctness of the discovered model, including the quality of the available data, the size of the sample obtained and the strength of the causal links to be discovered.

Having developed methods which, given large samples, discover causal models that are generally as good as or better than those discovered by TETRAD II [Wallace, Korb and Dai, 1996], we report here initial results on the robustness of the two methods when using small samples and in discovering weak links. In Section 2, we describe the sample size and weak link discovery problems. In Section 3, we give a brief analysis of the relationship between sample size, link strength and the discovery of causal links. Section 4 presents the test strategies. Section 5 provides the experimental results of the causal model discovery algorithms across a range of sample sizes and with various small path coefficients. In particular we compare the results of the MML induction system MML-CI (the MML Causal Inducer) [Wallace.Korb and Dai, 1996] with that of TETRAD II [Glymour *et* al, 1987; Schemes, 1994; Spirtes *et* al, 1993].

2 Robustness of Causal Discovery

Let $V = \{v_i : 1 \leq i \leq n\}$ (corresponding to random variables) be a set of nodes and $E \subset V \times V$ be a set of links, a causal model *M* is a directed acyclic graph $(\mathrm{DAG}) < V, E >$ together with numerical parameters reporting the strength of the connections, where $< x_i, x_j > \in E$ means that z, is a direct cause of X_j relative to *V*. Such directed acyclic graphs that are used to represent causal theories are variously called causal models, causal graphs, causal networks and belief networks [Cooper and Herskovits, 1991] and [Russel and Norvig, 1995]. A causal network gives a concise specification of the joint probability distribution [Pearl, 1988]. Each node in the causal network has a conditional probability table that quantifies the effects that the parents have on the node; linear causal networks (e.g., [Wright, 1934]) provide the same information under the assumption that each effect variable is a linear function of its parents, allowing the numerical parameters to be attached to causal links independently.

Recently, causal models (especially in the form of Bayesian nets) have been widely employed for the representation of the knowledge with uncertainty, including use in expert systems [Shafer, 1996]. In consequence, interest has grown in the learning of causal models as well. Various learning strategies have been developed. These methods include Spirtes et al.'s TETRAD I and II based upon significance tests for partial correlations, Pearl and Verma's approach [Pearl, 1988] and [Pearl and Verma, 1991] using conditional independencies, Hecker-

man's Bayesian approach [Heckerman, 1995] and [Heckerman *et al,* 1995]. In 1995, Madigan and York introduced Markov Chain Monte Carlo Model Composition *(MC^3)* [Mdigan *et al,* 1995] for approximate Bayesian model averaging (BMA) and recently further developed the *Gibbs* MC^3 and the *Augmented* MC'^3 algorithms [Mdigan *et al,* 1995] for the selection of Bayesian models. More recently Wallace *et al* developed the MML-CI [Wallace,Korb and Dai, 1996] based on MML induction, a Bayesian minimum encoding technique and Suzuki proposed a MDL (Minimum Description Length) principle based Bayesian Network learning algorithm using the branch and bound technique [Suzuki, 1996].

Here we examine the particular problem of the robustness of the two causal discovery algorithms which have been developed for inducing linear causal models, namely MML-CI and TETRAD II. In particular, we compare the models these algorithms produce when presented with varying sample sizes and samples generated from original causal structures with varying strengths of causal relationship. The robustness of the discovery technique in dealing with small samples is an important issue for machine learning, since autonomous, resource-constrained agents must be prepared to learn interactively with environments that will not tolerate unbounded sampling. We need to estimate the reliability of a derived model. Also, although we do not here report on large causal models, we would expect that problems with robustness with small samples for small models will manifest themselves also with large samples for large models, suggesting difficulty in scaling up a learning algorithm to cope with realistic examples of causal discovery. Here, the large model refers to the model with large number of links.

3 The Influence of Sample Size and Link Strength

For any learning technique which converges on the underlying probability distribution in a prediction task, the predictive accuracy will be sensitive to sample size, model complexity and the strength of the correlation between measured variables. In general, predictive accuracy of a recovered model will be a function of sample size, quality of the data and the ability of the learner [Dai, 1994]. In the discovery of causal models verisimilitude of the model discovered relative to the original model (and the probability distribution implied) will also be affected by sample size, model complexity and the strength of causal association between measured variables. For practical purposes, starting from similar prior domain information, better learning ability will reveal itself in *faster* convergence upon the underlying model, or, to put it the other way around, in the robustness of discovery given smaller sample sizes. Here we examine such robustness in MML-CI and TETRAD II.

The probability of discovering from sample data the existence of a particular causal link depends, in part, upon the strength of that causal link. In the case of a single causal path between two nodes being a single, direct causal link (in standardized models) the path coefficient is identical to the correlation between the two nodes, making the relation between sample size and detectability of the link plain. TETRAD II is sensitive to the strength of the causal relation quite directly: it determines whether a link is present or not by applying significance tests potentially to all orders of partial correlation, removing the effects of all subsets of V excluding the nodes under consideration. In consequence, ordinary concerns about the robustness of significance testing apply to TETRAD II — and for each link these concerns will apply not to a single significance test, but to a battery of significance tests.[1] Things are worse than ordinary for TETRAD II, however: because a high-order partial correlation estimate depends upon estimates of the marginal correlations for each pair of variables involved, the uncertainties associated with each estimate will accumulate, which results in high standard errors (variance) for high-order partial correlation estimates and in the need for very large samples to get significant results. The reliance on significance tests for high-order partial correlations suggests that TETRAD II will be unlikely recover the structure of a larger model without quite large samples available. In other words, the larger the order of such a significance test, the greater the sample size must be for an effect of constant strength to be detected. As a result, as the authors admit [Scheines, 1994], TETRAD II has a tendency to omit arcs for larger models even with fairly large sample sizes.

MML-CI does not depend upon a test as rigid as significance tests at a fixed level: it reports an arc whenever the presence of such an arc leads to a reduction in the message length for a joint encoding of the causal model and the sample data [Wallace,Korb and Dai, 1996]. That is, given a sample with m instances over n variables, the message length is calculated according to the following formulas:

$$L = L_{Model} + L_{(Data|Model)} \qquad \text{(i)}$$

This involves a trade-off between greater simplicity of the model (and commensurately higher prior probability) and greater accuracy in accommodating the given sample data (and so a higher likelihood for the model) via Shannon's definition of information.

In detail, the MML encoding of causal models and data is given in the following equations (see [Wallace,Korb and Dai, 1996] for a detailed explanation). We

[1] This is true even though TETRAD II takes steps to reduce the number of significance tests required per pair of nodes, in its "PC algorithm" [Scheines, 1994].

start by dividing the code for the model into two parts, corresponding to the causal structure and the numerical parameters:

$$L_{Model} = L^{(s)} + L^{(p)} \quad (2)$$

We use

$$L^{(s)} = \log n! + \frac{n(n-1)}{2} - \log M \quad (3)$$

which provides an efficient encoding for a directed acyclic graph, when M is a count df the linear extensions of the dag.

$$L^{(p)} = \sum_{i=1}^{n} [\frac{n_i}{2} \log 2\pi + n_i \log \alpha_i + \frac{1}{2\alpha_i^2 \sigma_i^2} \sum_{k=1}^{n_i} a_{ik}^2 + \frac{1}{2} \log(2m) + \frac{1}{2} \log |A|] \quad (4)$$

where: n_i is the number of arcs incident on the variable x_i; α_i is a hyper-parameter reflecting the expected strength of the causal links to node i (set to 1 in all experiments reported here); σ_i is the variance; a_{ij} are the parameters; and $A = (x_i \cdot x_j)$ is the $n \times n$ data matrix. To encode the data requires a message of size:

$$L_{(Data|Model)} = \sum_{i=1}^{n} [\frac{m}{2} \log 2\pi + m \log \sigma_i + \sum_{j=1}^{m} \frac{r_{ij}^2}{2\sigma_i^2}] \quad (5)$$

where r_{ij} is deviation of the data from the linear prediction.

In MML-CI's discovery the relation between sample size and.the strength of causal links remains, of course; but the possibility of MML-CI finding weaker links sooner seems intuitively more likely, because such links will be reported as soon as the improvement they afford in encoding the data overcomes the increased cost of reporting a somewhat more complex model.

4 Testing Strategy

To examine the influence of sample size on the discovery of causal models experimentally we chose six models varying in complexity: models 1 through 6 in Figure 1-. We used these models to generate sets of sample data of various sizes stochastically, which in turn were given as input to MML-CI and TETRAD II to determine what causal models would be discovered. In the case of TETRAD II default values were employed exclusively; no prior information about the temporal order of variables was provided to either algorithm.

The first model is simplest, having only one link and two variables. In this case, the path coefficient is exactly equal to the correlation between the two variables. This makes the existence of the causal link extremely easy to find (although not its direction). Model six is the most complex model, having five variables and seven

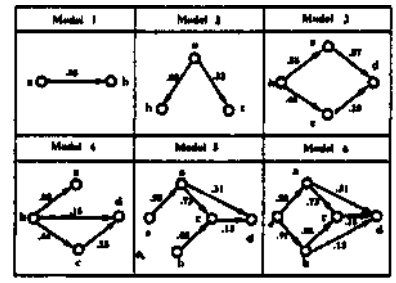

Figure 1: Six test models

arcs. Three of the models contain weak links with coefficients less than 0.1, namely models 3, 5 and 6. These six artificial models were manually designed for the following testing purposes: (1) The learning difficulties associated with model complexity in terms of the number of variables; (2) The learning difficulties associated with model complexity in terms of the number of arcs; (3) The learning difficulties associated with the strength of the links. In each case we generated data sets with 10, 50, 100, 200, 500, 1000, 2000 and 5000 instances. Then we ran both MML-CI and TETRAD II using all eight data sets for each of the six models.

In a second experiment we looked at the effect of link strength on the recovered model. In this case we used model 6 (above) with the strength of the causal arc $b \rightarrow c$ varied between 0.08 and 0.16, in each case generating the same range of sample sizes as above.

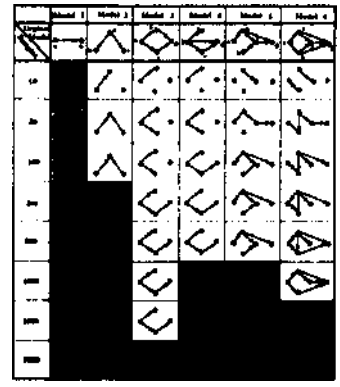

Figure 2: MML-CI Sample Size Test Results

5 Experimental Results and Analysis

Sample Size and Model Complexity In our experiments, we focus on linear models with Gaussian error and assume no hidden variables. We use TETRAD II default settings with a significant level of 0.05. The PC algorithm is the one applied on fully measured models with continuous variables. Figure 2 reports the mod-

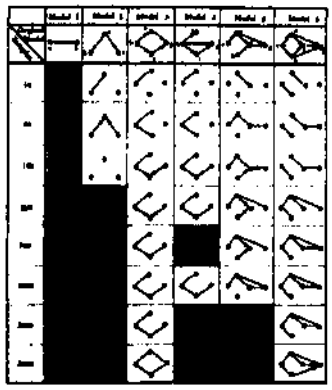

Figure 3: TETRAD II Sample Size Test Results

els discovered by MML-CJ from the 48 data sets, while Figure 3 reports those discovered by TETRAD II. The shading indicates for each model at what point the algorithm discovered the original model or a model statistically equivalent to the original. *Statistically equivalent* causal models are those which can be used to specify the same class of probability distributions over the variables (perhaps using distinct parameterizations). [Verma and Pearl, 1990] report a simple graphical criterion of equivalence which can be used to identify the statistically equivalent models in our figures: two causal models are statistically equivalent if and only if they have the same skeleton (undirected graph) and they have the same *v-structures* (nodes that are the children of two parents which are themselves non-adjacent). Such models cannot be distinguished on the basis of sample data alone [Chickering, 1995], so the discovery of one is as good as the discovery of another in this experiment.

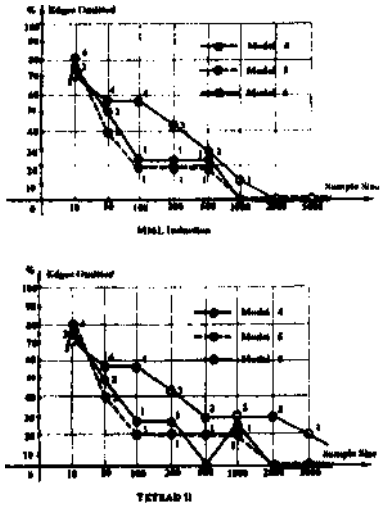

Figure 4; Comparison of Edges Omitted

For TETRAD II, in Figure 3, undirected arcs reflect the fact that TETRAD was unable to determine an arc orientation; for these arcs, either orientation is allowed by TETRAD, so long as the resulting graph is acyclic and so long as no new v-structures are introduced by selecting such an orientation. We counted the resulting TETRAD graph as satisfactory (and so appears shaded) if no such selection of arc orientations results in a causal model that is not statistically equivalent to the original model.

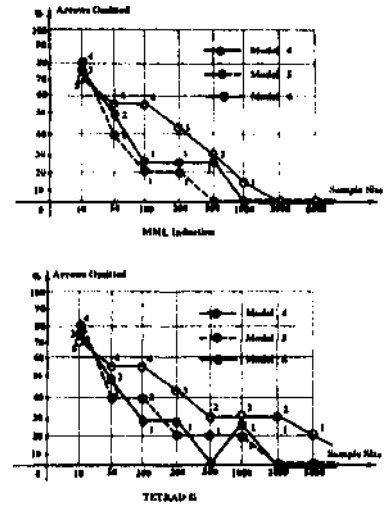

Figure 5: Comparison of Arrows Omitted

Although in this study we have not performed significance tests on the results (i.e., by generating large numbers of samples of each model for each sample size), the trend is fairly clear. For all of the models showing any complexity (i.e., for model 3 and above) MML-CI has found the correct model at smaller sample sizes than has TETRAD II. In the case of model 6 TETRAD II was unable to recover the weakest link even when supplied 5000 samples, while for model 3 TETRAD II found all the links but failed to discover the v-structure at node *d*.

Figures 4 and 5 compare MML-CI with TETRAD II in the manner used by Spirtes, et al. [Scheines, 1994]. Figure 4 graphs the percentage of edges of the original model which MML-CI and TETRAD II have failed to recover, by sample size. Figure 5 graphs the percentage of arc orientations missed by each program (but not counting cases where a graph with an incorrect arc orientation is statistically equivalent to the original model). Of course, both algorithms display the expected convergence towards zero errors — expected because TETRAD II is, in effect, a classical estimation technique whereas MML-CI is, in effect, a Bayesian estimation technique,

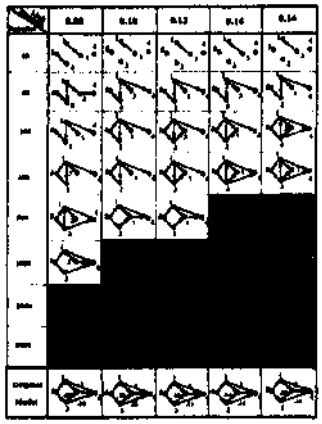

Figure 6: MML-CI Weak Link Discovery Results

and so both fall under the general convergence results established for the respective classes of statistical inference procedures. It remains of interest, however, that in all of these measures MML-CI tends to display a more rapid convergence towards the true model — which is to say it appears to be more robust when dealing with smaller sample sizes.

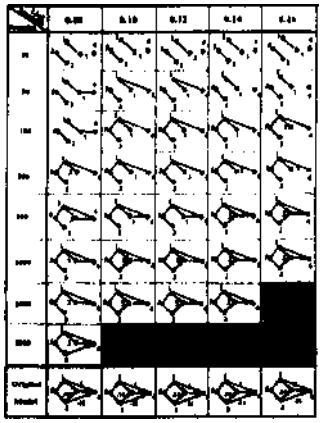

Figure 7: TETRAD II Weak Link Discovery Results

Sample Size and Weak Link Discovery Figure 6 illustrates the experimental results for MML-CI on model 6 when the causal link from *b* to *c* takes varying degrees of strength, in particular coefficients ranging from 0.08 to 0.16. Unsurprisingly, the results clearly reveal the fact that the weaker the association the larger the sample required to discover it. With the weakest coefficient of 0.08 in Figure 6, MML-CI does not discover the link until provided with 2000 samples. Whereas with a weakest link of 0.10 and 0.14, the system discovered the link once provided with a data set with the sample size of 500 and 100 respectively.

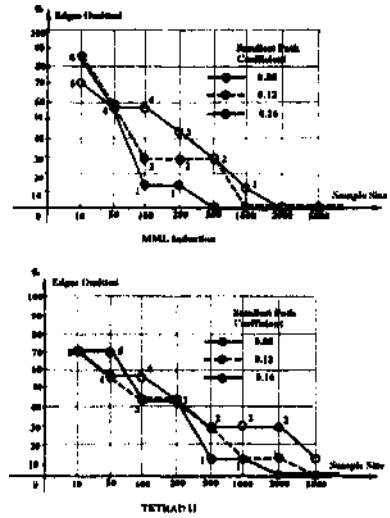

Figure 8: Comparison of Edges Omitted With Small Path Coefficients

Figure 7 illustrates like experimental results for TETRAD II. These results again show the inverse relationship between strength of causal relationship and the sample size required to discover it. Given coefficients above our original 0.8 TETRAD II was able to discover the link between 6 and *c* that it had missed before. It remains clear in all of the test cases that MML-CI recovers the original causal model with fewer samples than TETRAD 11. Figure 8 and Figure 9 report similar stories for the measures of arc omission and arrow omission.

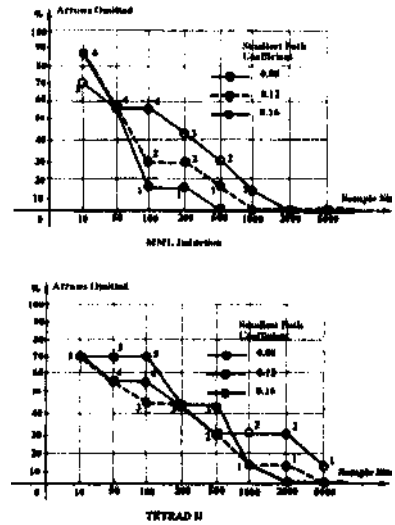

Figure 9: Comparison of Arrows Omitted With Small Path Coefficients

6 Conclusions

The following conclusions appear to be supported by our experimental results. (1) The theoretical difficulties of significance testing with robustness appear to be manifested in TETRAD II's inferior robustness with respect to sample size. This shows up, for example, in TETRAD IPs inability to recover the weaker links (with coefficients below 0.1) with smaller samples. (2) The problem of arc omission given small samples is particularly acute for TETRAD 11 (in comparison with MML-CI) as model complexity increases, as predicted by our analysis in §3. From the experimental results we also find that MML-CI shows promise not just in finding causal models that are as good as those discovered by TETRAD II in general, but given the constraints imposed by small samples or by weak causal links the models discovered appear to be characteristically superior to those discovered using the significance testing methods of TETRAD. This is likely to be an especially important feature of causal discovery when causal models become large, for TETRAD's method of examining partial correlations of all orders in such cases is both computationally expensive and lacking robustness.

Acknowledgement

This work was conducted with partial assistance from ARC grant .449330662.

References

[Cooper and Herskovits, 1991] G. F. Cooper and E. Herskovits. A Bayesian method for constructing Bayesian belief networks from databases. In *Proc. of 7th Conference on Uncertainty in Al.* Morgan Kaufmann, 1991.

[Chickering, 1995] David M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proc. of the 11th Conference on Uncertainty in Artificial Intelligence,* pages 87-98, 1995.

[Dai, 1994] Honghua Dai. Learning of forecasting rules from large noisy real meteorological data. *PhD. Dissertation, Department of Computer Science, RMIT,* 1994.

[Glymour *et al,* 1987] Clark Glymour, Richard Scheines, Peter Spirtes, and Kevin Kelly. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling.* Academic Press, San Diego, 1987.

(Heckerman, 1995] David Heckerman. A Bayesian Approach to Learning Causal Networks. *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence,* pages 285-295, 1995.

[Heckerman *et al,* 1995] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning,* 20(3): 197-243, 1995.

[Mdigan *et* al, 1995] David Madigan, Steen A. Andersson, Michael D. Perlman, and Chris T. Volinsky. Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. *NIPS 95 Workshop on Learning in Bayesian Networks and Other Graphical Models,* 1995.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, California, 1988.

[Pearl and Verma, 1991] Judea Pearl and T. S. Verma. A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference,* pages 441-452, San Mateo, California, April 22-25, 1991. Morgan Kaufmann Publishers.

[Russel and Norvig, 1995] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Mordern Approach.* Prentice Hall, Englewood Cliffs, New Jersey 07632, 1995.

[Spirtes *et al,* 1993] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer-Verlag, New York, Berlin, Heideberg, 1993.

[Shafer, 1996] Glen Shafer. *Probabilistic Expert Systems.* SIAM Press, 1996.

[Scheines, 1994] R. Scheines, P. Spirtes, C. Glymour, and C. Meek. *TETRAD II: tools for causal modeling.* Lawrence Erlbaum Associates, Inc., Publishers, 365 Broadway, Hillsdale, New Jersey 07642, 1994.

[Suzuki, 1996] Joe Suzuki. Learning bayesian belif networks based on the minimum description length principle: An efficient algorithm using the B & B techniques. In Lorenza Saitta, editor, *Proceedings of the 13th International Conference on Machine Learning,* pages 462-470, 340 Pine St., 6th Floor, San Francisco, July, 1996. Morgan Kaufmann.

[Verma and Pearl, 1990] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence,* pages 220-227, Boston, MA, 1990. Morgan Kaufmann.

[Wallace,Korb and Dai, 1996] Chris Wallace, Kevin Korb, and Honghua Dai. Causal discovery via MML. In *Proceedings of the 13th International Conference on Machine Learning (ICML'96),* pages 516-524, 1996.

[Wright, 1934] Sewall Wright. The method of path coefficients. *Annals of Mathematical Statistics,* 5:161-215, 1934.