Proposal for the Non-parametric Test of Conditional Independence

Chu, Tianjiao

March 19, 2001

Abstract

Causal graphical models has been an active research area for the past decade, mostly because it provides a way to derive causal information from purely observational data. Current algorithms for causal inference require conditional independence relations as input. However, the only available tests for conditional independence are designed for data that are either discrete, or generated from linear models. To apply causal inference algorithms to general continuous data, we will need non-parametric conditional independence tests. In this project we will study two new approaches to the non-parametric test of conditional independence relationship for continuous data. The first approach is the regression approach, with the additive assumption. The second approach is the density estimation approach. The two approaches will be evaluated with samples generated from additive models. The type one and type two errors of these two approaches will be compared. We do not know yet which one will perform better, but our guess is that the first approach will work only on a smaller set of cases, while the second approach will work on a larger set of cases. However, it is possible to identify the cases where the first approach works, and the performance of the first approach will be good on these cases. The performance of the second approach on those smaller set of cases might not be as well as that of the first approach.

Our simulation study will focus on the conditional independence test among a small number (3 or 4) of random variables, because in practice, both approach will be reliable only when they are used to test the conditional independence relationship among 3 or 4 random variables. However, it is known that even low dimensional conditional independence information will provide information about the causal structure of the data.

1 Introduction

Causal information, or even partial causal information, can help decision making. For example, the knowledge that lung cancer and smoking are positively correlated does not tell us what we should do to prevent lung cancer. However, if we know that smoking is a cause of lung cancer, we know that we can reduce the risk of lung cancer by the control of smoking.

Traditionally, causal information is derived from controlled experiments. However, in many cases, controlled experiments are too expensive, or practically infeasible, or ethically improper. Nevertheless, in these cases, observational data are often available. In the past decade, work by Pearl, Spirtes, Glymour, Scheines, and other people,[1] has shown that we can make inference of a causal network, or a causal pattern, from purely observational data, provided we could retrieve conditional independence information among the random variables from these data. [2] For example, if we have three variables X, Y, and Z. Suppose we know that both X and Z are direct causes of Y, and there are no other common causes of any two of these three variables. Then we know that X and Z are independent, and conditional dependent give Y. Conversely, the set of conditional independence relationships among the variables all impose some constraints on the possible causal relations among the variables.

In order to get causal information from purely observational data, we must be able to test conditional independence relationships among the random variables. In this project, we propose two approaches to the conditional independence tests: the regression approach and the density estimation approach. Both approaches can be applied to a large class of observational data consisting of continuous variables. In particular, the regression approach will allow that the expectation of a random variable conditional on its direct causes can be a linear combination of arbitrary smooth univariate functions of its direct causes. The density approach will put virtually no constraint on the continuous variables. Because of the lack of asymptotic results on their behavior, we will use simulation to study their performance.

In the following sections, we will first discuss the prior work on test of independence and conditional independence, and their limitations, and then present the algorithm of the two new proposals and our simulation plan.

2 Prior Work

Conditional independence test for a set of discrete random variables is trivial. The χ^2 test of goodness-of-fit can be used as conditional independence test.[3] For linear models with a normal distribution [4], it is known that a certain function of the sample partial correlation could be

[1] Pearl (1988), Spirtes et al (2000).

[2] The basic idea is that we can use directed acyclic graph (DAG) to represent causal relations among a set of random variables. Each vertex will represent a random variable, and there is a directed edge from vertex X to vertex Y if and only if the random variable corresponding X is a direct cause of the random variable corresponding Y. If we further impose two reasonable conditions, i.e., the **Markov Condition** and the **Faithfulness Condition** (see Spirtes et al (2000)), then the conditional relationships among the random variables impose some constraints on the conditional relationships among the random variables.

[3] Chapter 4, Conover (1980).

[4] A linear model assumes that each random variable in the model is a linear combination of its direct causes, plus some independence random noise.

used as a statistic for conditional independence test. [5] Alternatively, we could use Fisher's z statistic, a simple function of which has a standard normal distribution.

However, we often need to know the conditional independence relations among a set of continuous variables where we have no reason to assume that the data came from a linear model. Indeed, given that we do not even know which variable causes which one, how can we say that the functional relations among the variables are all linear? In these cases, non-parametric conditional independence tests for continuous data should be applied.

There is not much literature about non-parametric conditional independence test. [6] The only paper on this topic I have ever found is Linton et al (1999). However, the approach proposed in that paper is designed to test a null hypothesis that is much stronger than the conditional independence hypothesis, and hence in general will have very high rate of type 1 error. That is, it is highly likely that the test will reject the null hypothesis (conditional independence) even if it holds. There is a large set of methods for non-parametric independence test between two random variables, such as the Spearman's ρ test and Kendall's τ test[7], Kallenberg et al 1999, Gieser et al 1999, and Kano 1992. However, these methods cannot be extended to test conditional independence, unless the conditioning variable is discrete. The reason is simple: to show that X and Y are conditionally independent given Z, we have to show that X and Y are independent conditional on every possible value of Z. Obviously, this is impossible when Z is continuous.

Friedman et al (1996) uses the discretization method to learn a Bayes network from continuous data. This suggested that one possible approach to the conditional independence test is to convert the continuous data into discrete data, and then apply the χ^2 test to the discretized data. I am not aware of any report about this approach. The problem with this approach is that the conditional independence relation among the continuous variables does not imply conditional independence among the discretized variables. Conversely, conditional independence among discretized variables does not imply conditional independence among continuous variables either. [8] We could improve the approximation of the continuous distribution by increasing the number of bins for discretization, but the conditional independence test for discrete data is not reliable when the number of observations per cell is small.

3 Our Proposal

In this project, we are going to study two approaches to the conditional independence test that do not require the linear assumption, nor do they require the data to be discretized. This is a continuation of my master's project.

[5] §3.4, Jobson (1991).

[6] This lack of literature might be partly attributed the fact that causal inference study is relatively a new area.

[7] Chapter 5, Conover (1980).

[8] The measure theoretic definition of conditional independence is give as:

Let $X_1, \cdots, X_n$ be n random variables on a probability space $(\Omega, \mathcal{F}, P)$. Then $\{X_1, \cdots, X_n\}$ are conditional independent given Y if $\forall A_i \in \sigma(X_i)$, $P\left(\bigcap_{i=1}^n A_i \middle| Y\right) = \prod_{i=1}^n P(A_i|Y)$ amost surely with resepct to P_Y, where $\sigma(X_i)$ is the σ-field on Ω generated by X_i, and P_Y is the restriction of P to the $\sigma(Y)$, the σ-field generated by Y.

However, this does not imply $\forall A_i \in \sigma(X_i)\ \forall B \in \sigma(Y)$, $P\left(\bigcap_{i=1}^n A_i \middle| B\right) = \prod_{i=1}^n P(A_i|B)$.

Actually, if we let $B = \Omega$, the latter formula will imply that $X_1, \cdots, X_n$ are independent!

3.1 Regression Approach

In my master's project, I did a simulation study of the regression approach to the conditional independence test with the additive assumption. The rationale behind the regression approach is that, in general, the variance of X is greater than or equal to the expected variance of X given Y, that is, $\mathrm{E}[\mathrm{Var}(X|Y)] \leq \mathrm{Var}[X]$. However, if two variables X and Y are independent, then we are not going to predict X better even if we know Y. That is, $\mathrm{E}[\mathrm{Var}(X|Y)] = \mathrm{Var}[X]$.

Note that although independence implies that $\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|Y)]$, the converse does not hold. [9] However, in the practice of causal inference, we assume that random variables are causally connected in such a way: If Y is the parent of X, then $\mathrm{E}[X|Y] = f(Y)$, where f is some arbitrary non-constant smooth function. Therefore, it will be rarely the case that X and Y are dependent and $\mathrm{E}[X|Y] = \mathrm{E}[X]$.

The general algorithm of regression method is:

1 Estimate $\mathrm{E}[X|\boldsymbol{Z}]$ and $\mathrm{E}[X|Y, \boldsymbol{Z}]$ by regress X on $\boldsymbol{Z}$ and $\{Y, \boldsymbol{Z}\}$ respectively.

2 Score the fits of X on $\boldsymbol{Z}$ and on $\{Y, \boldsymbol{Z}\}$ using some score function, e.g., BIC. [10]

3 $X \perp\!\!\!\perp Y|\boldsymbol{Z}$ if and only if the fit of X on $\boldsymbol{Z}$ is better than that of X on $\{Y, \boldsymbol{Z}\}$.[11]

There are many different regression methods we can use to estimate $\mathrm{E}[X|\boldsymbol{Z}]$ and $\mathrm{E}[X|Y, \boldsymbol{Z}]$, such as surface smoother, project pursuit, etc. However, most of these methods are subject to the curse of dimensionality, i.e., their performance will be unreliable for high dimension data set due to the scarce of data points in a high dimensional cube. Thus, we make a further assumption, i.e., we assume that the data were generated from an additive model. In an additive model, the response variable is assumed to be the sum of arbitrary univariate functions of each predictor, plus an error term. Let X be the response variable, and $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ be the direct causes of X, then the additive assumption implies that:

$X = \sum_{i=1}^{n} f_i(Y_i) + \epsilon$,

where f_i's are arbitrary smooth functions, and ϵ is independent of (Y, Z). Note that without the additive assumption, we would have assumed that $X = g(Y_1, \cdots, Y_n) + \epsilon$.

The additive assumption is general enough for many practical uses. For examples, it is widely assumed in semi-parametric models of economic data (see Gozalo (2001)). It allows us to use the additive model regression method. This method avoids the problem of dimensionality, because it regresses the response variable X on Y and Z separately.[12] Moreover, our simulation studies showed that small violations of additive assumption do not affect the power of our tests.

The simulation study in my master's project showed that, on the one hand, the accuracy of the additive regression method is not satisfactory; on the other hand, some information

[9] One example where $\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|Y)]$ but X and Y are dependent is the case where $Y > 0$ and $X \sim N(0, Y)$. Another example is the case where $X \sim N(0, \sigma_1^2)$, $\epsilon \sim N(0, \sigma_2^2)$, and $Y = X^2 + \epsilon$, for $\mathrm{E}[X] = \mathrm{E}[X|Y] = 0$.

[10] BIC, the Bayesian information criterion, also called the Schwartz information criterion, in first proposed in Schwartz (1978). It penalizes the log-likelihood of a model by $\frac{\log N * df}{2}$, where N is the sample size, and df is the number of parameters.

[11] The symbol $\perp\!\!\!\perp$ is used to represent independence relation. $X \perp\!\!\!\perp Y|\boldsymbol{Z}$ means that X and Y are independence conditional on $\boldsymbol{Z}$.

[12] For the algorithm of the (generalized) additive model regression, please see Hastie et al (1990).

provided by the regression can be used to predict the reliability of the conditional independence test. Consider a set of variables $\boldsymbol{V} = \{X, Y, Z_1, \cdots, Z_k, U_1, \cdots, U_m\}$. Suppose we want to test whether $X \perp\!\!\!\perp Y | (Z_1, \cdots, Z_k)$. Let *null deviance* be the sample variance of X, *noise deviance* be the residual sum square of regressing X on $(Y, Z_1, \cdots, Z_k, U_1, \cdots, U_m)$, *independent deviance* be the residual sum square of regressing X on $Z_1, \cdots, Z_k$, and *dependent deviance* be the residual sum square of regressing X on $Y, Z_1, \cdots, Z_k$. We believe that the 3 ratios, i.e., $\frac{\textit{noise deviance}}{\textit{null deviance}}$, $\frac{\textit{independent deviance}}{\textit{null deviance}}$, and $\frac{\textit{dependent deviance}}{\textit{null deviance}}$, can help us to estimate the sensitivity of the conditional independence test to small perturbation in the additive regression model, hence can be used to estimate the reliability of the conditional independence test. We proposed the following heuristic rules, assuming we use the BIC score:

1. If $\frac{\textit{noise deviance}}{\textit{null deviance}}$ is very small, and $\frac{\textit{independent deviance}}{\textit{null deviance}}$ is not very close to 1, then if the result is dependence, it is a **weak dependence**, and if the result is independence, it is a **strong independence**.

2. If $\frac{\textit{noise deviance}}{\textit{null deviance}}$ is relatively large, and $\frac{\textit{independent deviance}}{\textit{null deviance}}$ is not very close to 1, then if the result is dependence, it is a **strong dependence**, and if the result is independence, it is a **weak independence**.

3. If both $\frac{\textit{independent deviance}}{\textit{null deviance}}$ and $\frac{\textit{dependent deviance}}{\textit{null deviance}}$ are very close to 1, then if the result is dependence, it is a **weak dependence**, and if the result is independence, it is a **weak independence**.

What we are going to do in this project are, first, to find more heuristic rules to determine the reliability of a conditional independence test, second, find out whether they hold in general, and third, if these rules do hold, try to quantify them and give some justification for these rules. More precisely, we will first do the additive regression, and then train a decision tree, call it the *test tree*, with some the information we can get from the regression result, including BIC score, etc., to predict whether a certain conditional independence relation holds. Then we use some other information, including the 3 ratios we mentioned above, to train another decision tree, call it the *meta tree* to estimate how confident we are about the results given by the the first tree. Compared to method we studied in the master's project, the combination these two trees with the additive regression, we believe, will be a more promising approach. We expect this new approach will give both higher accuracy, and better knowledge about when the test result is reliable.

We plan to use Splus as the simulation environment. Splus has implemented the back-fitting algorithm for additive regression proposed in Hastie et al (1990). We will use spline smoother as the univariate regression module for the back-fitting algorithm.

3.2 Density Estimation Approach

The density estimation method is conceptually straightforward. The basic idea is to estimate marginal distributions and joint distribution, and see how close a certain function of the marginal distributions is to the joint distribution.

The main problem of the density estimation method is that it is subject to the curse of dimensionality. If the number of variables is n, we need to estimating a joint distribution in an n-space by estimate locally weighted empirical distributions of the sample points within a neighborhood, say with radius r. For the estimation to be reliable, there must be sufficient sample points in this neighborhood. However, as the dimension n increases, the sample size must also increase exponentially with n to maker sure that a neighborhood of the same radius r contains enough sample points. In other words, to estimate density for high dimensional data, the sample size must be very large.

It is precisely because of the concern over the curse of dimensionality that we chose the regression method with additive assumption to test conditionally independence. However, the preliminary simulation study of the regression method shows that the latter method is not reliable for high dimensional data, which means that application of regression method should be restricted to low dimensional tests. Moreover, Spirtes (2001) shows that low dimension conditional independence relations can still help us to infer cause patterns from observational data set. Therefore, there is now no reason to prefer the regression method over the density estimation method.

Suppose we want to test whether X is independent of Y given Z, we estimate the pdf $\hat{f}_{x,z}(x,z)$ of X,Z, the pdf $\hat{f}_{y,z}(y,z)$ of Y,Z, the pdf $f_z(z)$ of Z, and the pdf $f_{x,y,z}(x,y|z)$ of X,Y,Z. Then we can test the null hypothesis that X is independent of Y given Z using a distance between $f_{x,y,z}(x,y,z)$ and $\dfrac{f_{x,z}(x,z)f_{y,z}(y,z)}{f_z(z)}$ as the test statistic.

Note that the distance D between the two densities $f_{x,y,z}(x,y,z)$ and $\dfrac{f_{x,y}(x,y)f_{y,z}(y,z)}{f_z(z)}$ could be any distance measure.

Let $D = \int \delta(x,y,z)\,dx\,dy\,dz$, then one choice of δ is:

$$\delta(x,y,z) = \left(\sqrt{\hat{f}_{x,y,z}(x,y,z)} - \sqrt{\frac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}}\right)^2 \tag{1}$$

which is the kernel of Hellinger distance.

We may also want to use

$$\delta(x,y,z) = \left|\hat{f}_{x,y,z}(x,y,z) - \frac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}\right| \tag{2}$$

which is the kernel of L_1 distance, or

$$\delta(x,y,z) = \left(\hat{f}_{x,y,z}(x,y,z) - \frac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}\right)^2 \tag{3}$$

which is the kernel of L_2 distance.

The algorithm will be:

1 Estimate $f_{x,y,z}(x,y,z)$, $f_{x,y}(x,y)$, $f_{y,z}(y,z)$, and $f_z(z)$,and compute D.

2 Generate N set of bootstrap samples $S_1, \cdots, S_N$ from the distribution $\dfrac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}$. Compute D_i for each set of data S_i, and approximate the CDF F_d of D_i.

3 Compute $F_d(D)$, and accept the null hypothesis at a significant level α if $F_d(D) \leq \alpha$.

This method will be slow, because we need to do multi-dimensional density estimation N times. Also, because the multi-dimensional density estimation in general is not as reliable as univariate regression, we do not expect that the density estimation method would outperform the regression method in the those cases where the regression method is supposed to be reliable, that is, the cases where the meta tree feels confident. However, we believe that the density method could be applied to a broader class of data set, including the cases where we have mixture of continuous and discrete variables.

We plan to use Splus as the simulation study environment, and use the kernel method to estimate the density functions. The density estimation for one dimensional data is available. We will need to implement density estimation algorithm for 2 and 3 dimensional data.

3.3 Evaluation Method

We will apply the two algorithms to a set of synthetic data to estimate their rates of two types of errors. The type 1 errors are the cases where the null hypothesis holds, i.e., the conditional independence relation exists, and the test rejects the null hypothesis. The type 2 errors are the cases where the null hypothesis is false, i.e., the conditional independence relation does not hold, and the test accepts the null hypothesis. The data will be generated from five different causal graphical models.

4 Work Plan

My work plan for the following 8 weeks (Feb 5 — April 1) is:

Week 1 Do research on the density estimation algorithm.

Week 2 Incorporate the reviews to the project, continue research on the density estimation algorithm.

Week 3 Implement the regression algorithm.

Week 4 Finish the regression algorithm, begin implement the density estimation algorithm.

Week 5 Continue the implementing of density estimation algorithm.

Week 6 Finish the density estimation algorithm, begin to evaluate the implemented algorithms on simulated data.

Week 7 Finish the evaluation of the algorithms.

Week 8 Write final report.

5 Summary

As stated before, the past decade had seen a growing interest in the study of the causal inference from purely observational data. Many algorithms were developed to handle various possible causal structures. All the causal inference algorithms assume the availability of complete conditional independence information.

However, the current available methods for conditional independence test can only be applied to purely discrete data or purely continuous data generated from linear models. As a matter of fact, after several weeks of intensive search, I could only find one report of the study of the non-parametric conditional independence test, i.e., Linton et al (1999), which, as I said before, has a high rate of type 1 error. This is rather surprising given the fact that in the past decade causal inference, which is entirely based on the conditional independence information, has been an active area of the research. The lack of general method for conditional independence test has so far confine the application of causal inference to a small set of cases. It is our hope that this project will allow users to use the causal inference tools to deal with real world data sets, which are unlikely generated from some linear models, or purely discrete.

Reference

Akaike, H. (1973) "Information Theory and an Extension of the Maximum Likelihood Principle", in *Second International Symposium on Information Theory*, edited by B. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267-81.

Conover, W. (1980) *Practical Nonparametric Statistics* 2nd edition, New York: John Wiley & Sons.

Friedman, N., Goldszmidt, F. (1996) "Discretizing Continuous Attributes While Learning Bayesian Networks", in *Thirteenth International Conference on Machine Learning.*

George, E. (2000) "The Variable Selection Problem", unpublished paper.

Gieser, P. W., Randles, R. H.A (1997) "A Nonparametric Test of Independence Between Two Vectors", *Journal of the American Statistical Association*, Vol. 92, pp. 561-567.

Gozalo, P., Linton, O. (2001) "Testing Additivity in Generalized Nonparametric Regression Models with Estimated Parameters," forthcoming in *Journal of Econometrics.*

Hastie, T. Tibshirani, R. (1990) *Generalized Additive Models*, New York: Chapman and Hall.

Jobson, J. (1991) *Applied Multivariate Data Analysis Vol. I*, New York: Springer-Verlag.

Kallenberg, W., Ledwina, T. (1999) "Data-Driven Rank Tests for Independence", *Journal of the American Statistical Association*, Vol. 94, pp. 285-301.

Kano, Y. (1992) "Robust Statistics for Test-of-independence and Related Structural Models", *Statistics & Probability Letters*, Vol. 15, pp.21-26.

Linton, O., Gozalo, P. (1999) "Conditional Independence Restrictions: Testing and Estimation", *Cowles Foundation Discussion Papers*, No. 1140.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligence Systems*, San Mateo, Morgan Kaufmann.

Richardon, T. (1996) "A Discovery Algorithm for Directed Cyclic Graphs", in *Proceedings of the 12th Conference of Uncertainty in AI*, Portland, OR, Morgan Kaufmann: pp.454-461.

Schwartz, G. (1978) "Estimating the Dimension of a Model", in *Annals of Statistics Vol. 6*, pp.461-464.

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

Spirtes, P. (2001) "An Anytime Algorithm for Causal Inference", *Artificial Intelligence and Statistics 2001.*

Ye, J. (1998) "On Measuring and Correcting the Effects of Data Mining and Model Selection", *Journal of the American Statistical Association Vol. 93* pp.120-131.