Nonparametric Test of Conditional Independence

Tianjiao Chu TCHU@ANDREW.CMU.EDU
Department of Philosophy and CALD

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. In this paper, we studies two nonparametric conditional independence testing algorithms that could applied to a much broader class of continuous data: the additive regression approach and the density estimation approach. The simulation results showed that the additive approach performed well for an identifiable sets of data. The density estimation approach, while theoretically should work on a larger set of cases, could not provide reliable results.

1. Introduction

1.1 Causal Inference and conditional independence

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 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, consider three variables X, Y, and Z. If both X and Z are direct causes of Y, and there are no other common causes of any two of these three variables, then X and Z must be independent, and conditional dependent give Y. Conversely, the set of conditional independence relationships among the variables also impose some constraints on the possible causal relations among the variables. Hence we can infer some causal pattern from the information about the conditional independence relations.

1.2 Why do we need nonparametric tests

In order to get causal information from purely observational data, we must be able to test conditional independence relationships. Some conditional independence test algorithms have been developed and widely used for purely discrete data, or continuous data that are assumed generated from linear models. For example, if the data set is purely discrete, we could use the popular χ^2 statistic to test whether certain conditional independence relations hold. For the continuous data sets that are believed to be generated from linear models, we could either use the Fisher's z statistic, or the partial correlation, as the test statistic.

However, we often need to know the conditional independence relations among a set of continuous variables where we do not believe that all the functional relations among these variables are linear. For example,

[1] Pearl (1988), Spirtes, Glymour, & Scheines (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.

many economists believe that the log of hourly wage rate is a nonlinear function of the annual hours worked (Biddle & Zakkin, 1989). Furthermore, there are many cases where we know very little about the functional relations among the variables, and hence cannot assume them to be linear. In these cases, nonparametric conditional independence tests are required.

1.3 Summary of our approach

In this paper, we propose two approaches to the nonparametric conditional independence tests: the regression approach and the density estimation approach. To illustrate these two approaches, suppose we have 3 random variables X, Y, and Z, and would like to know whether X is independent of Y given Z. The basic idea of the regression approach is to compare the regression of X on Z and the regression of X on (Y, Z), and compare the residual sums of squares of these two regressions. [3] If the regression of X on Z has similar residual sum of squares as the regression of X on (Y, Z), we would say that X is independent of Y given Z, because with the knowledge of Z, knowing Y does not help us predict X. Otherwise, we would say that X and Y are dependent given Z. [4]

The density estimation approach assumes that X, Y, and Z has a joint density $f_{X,Y,Z}(X, Y, Z)$. Then X and Y are independent conditional on Z if and only if the joint density of X and Y given Z, $f_{X,Y|Z}(x, y|z)$, is equal to the product of the density of X given Z, $f_{X|Z}(x|z)$, and the density of Y given Z, $f_{Y|Z}(y|z)$. Therefore, to test whether X and Y are independent given Z, we simply estimate $f_{X,Y|Z}(x, y|z)$, and compare it with the estimation of $f_{X|Z}(x|z)f_{Y|Z}(y|z)$. If they are similar, we say X and Y are independent given Z. Otherwise, X and Y are dependent given Z.

1.4 Key results

The simulation study of the above two approaches shows that the regression approach has a reasonable performance, while the density estimation approach does not work well.

In particular, although the regression approach does not perform well in all cases, there are a large number of cases where the regression approach gives almost perfect results. More importantly, for each conditional independence test using the regression approach, we can predict very well if the result is reliable. On the other hand, even with relatively large sample size (2000), the rate of type I error for the density estimation approach is 40%, and the rate of type II error is 27%.

2. Approach

Suppose we have three continuous random variables X, Y, and Z. To test the null hypothesis that X and Y are independent given Z, we could use either the regression approach, or the density estimation approach. [5]

2.1 Regression Approach

The general algorithm for the regression approach is:

1 Regress X on Z and on (Y, Z) respectively.

2 Score the fits of the regressions of X on Z and on (Y, Z) using some score function, e.g., BIC. [6]

3 $X \perp\!\!\!\perp Y|Z$ if and only if the fit of the regression of X on Z is better than that of X on $\{Y, Z\}$.[7]

2.1.1 Regression and conditional independence

The rationale behind the regression approach is that if two variables X and Y are independent given a third variable Z, then knowing Y does not help our prediction of X if we already know Z. Therefore, the expected value of the variance of X given Z is the same as the expected value of the variance of X given Y and Z, if X and Y are independent given Z. (Of course here we assume that X^2 is integrable.)

More precisely, it can be shown that, the expected value of the variance of X given Z is greater than or equal to the expected value of variance of X given Y and Z. That is, $\mathrm{E}[\mathrm{Var}(X|Y, Z)] \leq \mathrm{E}[\mathrm{Var}(X|Z)]$:

$$\mathrm{E}[\mathrm{Var}(X|Y, Z)] = \mathrm{E}[\mathrm{E}[X^2|Y, Z] - \mathrm{E}[X|Y, Z]^2]$$

[3] Actually we also need to use some penalty term to penalize the more complex model. In this case, we need to penalize the regression of X on Y and Z, which has one more predictor than the regression of X on Z.

[4] Note that the regression of X on (Y, Z) should always has a residual sum of squares less than or equal to the regression of X on Z.

[5] Throughout this paper, we shall choose the hypothesis that the conditional independence relation holds as the null hypothesis.

[6] 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.

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

$= \mathrm{E}[X^2] - \mathrm{E}[\mathrm{E}[X|Y,Z]^2] \leq \mathrm{E}[X^2] - \mathrm{E}[\mathrm{E}[X|Z]^2]$ [8]

$= \mathrm{E}[\mathrm{E}[X^2|Z]] - \mathrm{E}[X|Z]^2] = \mathrm{E}[\mathrm{Var}(X|Z)]$

Note that although conditional independence implies that $\mathrm{E}[\mathrm{Var}(X|Z)] = \mathrm{E}[\mathrm{Var}(X|Y,Z)]$, the converse does not hold.

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

In practice, we do not need to worry about these situations. Usually, we assume that random variables are causally connected in such a way: If Y and Z are the parents of X, then $\mathrm{E}[X|Y,Z] = f(Y,Z)$, where f is some arbitrary non-constant smooth function. This essentially excludes the cases similar to the first example. For the cases similar to the second example, we could instead test whether Y is independent of X given Z by comparing the regression of Y on Z with the regression of Y on (X, Z).

2.1.2 Additive models

There are many different regression methods we can use in our conditional independence testing algorithm, 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 sets due to the scarce of data points in a high dimensional cube. To mitigate this problem, we make a further assumption: we assume that the data sets were generated from additive models. In an additive model, a response variable is assumed to be the sum of arbitrary univariate functions of each predictor, plus an error term. If X is a response variable, and $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ is the direct causes of X, then the additive assumption says that:

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

where the 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 example, it is widely used in semi-parametric models of economic data (see Gozalo (2001)). The advantage of the additive model assumption is that 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 (Hastie & Tibshirani, 1990). In this paper, we shall use the additive regression method in our nonparametric conditional independence testing algorithm.

2.1.3 Reliability of the regression approach

An ideal algorithm will give excellent results for almost all data sets. However, as the simulation study shows, our regression approach is not such an algorithm. Therefore, it is extremely important to know, for each test, if the test result is reliable. [9] Otherwise the algorithm would be useless.

Consider a set of variables $\boldsymbol{V} = \{X, Y, Z, U_1, \cdots, U_m\}$. Suppose we want to test the null hypothesis $X \perp\!\!\!\perp Y | Z$. Using regression approach, we can get the following information:

- RSS_X: residual sum of squares of regressing X on 1;
- σ_2: estimated variance of the error term associated with X, (it is proportional to $RSS_{X|\cdot}$, the residual sum of squares of regressing X on $(Y, Z, U_1, \cdots, U_m)$);
- $RSS_{X|Z}$: residual sum of squares of regressing X on Z;
- $RSS_{X|Y,Z}$: residual sum of squares of regressing X on (Y, Z).

Suppose we use BIC as the score function. Then the score for the regression of X on Z is:

$$BIC_{X|Z} = \frac{RSS_{X|Z}}{\sigma^2} + \log N * df_{X|Z}$$

The score for the regression of X on (Y, Z) is:

$$BIC_{X|Y,Z} = \frac{RSS_{X|Y,Z}}{\sigma^2} + \log N * df_{X|Y,Z}$$

where N is the sample size, and $df_{X|Z}$ and $df_{X|Y,Z}$ are equivalent degrees of freedom, a kind of model complexity measurements, of the two regressions respectively.

The regressing approach basically says that X is independent of Y given Z if and only if $BIC_{X|Z} <$

[8] Note that $\mathrm{E}[\mathrm{E}[X|Y,Z]^2] \geq \mathrm{E}[\mathrm{E}[X|Z]^2]$, because:
$\mathrm{E}[\mathrm{E}[X|Y,Z]^2 - \mathrm{E}[X|Z]^2]$
$= \mathrm{E}[\mathrm{E}[X|Y,Z]^2 - \mathrm{E}[X|Z]^2|Z]]$
$= \mathrm{E}[\mathrm{E}[X|Y,Z]^2|Z] - \mathrm{E}[X|Z]^2]]$
$= \mathrm{E}[\mathrm{E}[X|Y,Z]^2|Z] - 2\mathrm{E}[\mathrm{E}[X|Y,Z]|Z]\mathrm{E}[X|Z] + \mathrm{E}[X|Z]^2]]$
$= \mathrm{E}[\mathrm{E}[X|Y,Z]^2|Z] - 2\mathrm{E}[\mathrm{E}[X|Y,Z]\mathrm{E}[X|Z]|Z] + \mathrm{E}[X|Z]^2]]$
$= \mathrm{E}[\mathrm{E}[(\mathrm{E}[X|Y,Z] - \mathrm{E}[X|Z])^2|Z]] \geq 0$

[9] Note that reliability is different from accuracy. An unreliable test result is as likely to be true as to be false.

$BIC_{X|Y,Z}$. Obviously only a small part of information obtained from the regressions is used to test the conditional independence hypothesis. Therefore we could use the same information to predict the reliability of the conditional independence test.[10] Suppose we have a set of N conditional independence tests, and we know whether these tests are correct, then we can use the following algorithm to learn to predict the reliability of new conditional independence tests.

1. Put the N tests into two sets. The first set consists of those tests where the null hypothesis was accepted. The second set consists of the tests where the null hypothesis was rejected.
2. For each set of tests, use the following procedure to generate two fitted additive models.
   - Compute $\frac{RSS_{X|Z}}{RSS_X}$, $\frac{RSS_{X|Y,Z}}{RSS_X}$, and $\frac{RSS_{X|\cdot}}{RSS_X}$ for each test.
   - Create a vector C such that $C[i] = 1$ if the i^{th} test is correct, and $C[i] = 0$ otherwise.
   - Use the additive regression method to regress C on $\left(\frac{RSS_{X|Z}}{RSS_X}, \frac{RSS_{X|Y,Z}}{RSS_X}, \frac{RSS_{X|\cdot}}{RSS_X}\right)$.
3. If a new conditional independence test accepts the null hypothesis, use the additive model learned from the first set of tests to predict its reliability. If the new test rejects the null hypothesis, use the model learned from the second set.

2.2 Density Estimation Approach

The density estimation approach to conditional independence testing is conceptually straightforward. Suppose we want to test the null hypothesis $X \perp\!\!\!\perp Y|Z$. We first estimate the joint density $\hat{f}_{x,z}(x,z)$ of (X,Z), the joint density $\hat{f}_{y,z}(y,z)$ of (Y,Z), the marginal density $f_z(z)$ of Z, and the joint density $f_{x,y,z}(x,y,z)$ of (X,Y,Z). Then we use some distance D between $f_{x,y,z}(x,y,z)$ and $\frac{f_{x,z}(x,z)f_{y,z}(y,z)}{f_z(z)}$ as the test statistic.

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

Let $D = \int \delta(x,y,z)\,dx\,dy\,dz$, our choice of δ is the kernel of Hellinger distance:

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

For a sample S, 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, $D(S)$, the D statistic for the sample S.

2 Generate B bootstrap samples $S_1, \cdots, S_B$ from the distribution $\frac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}$. Compute $D(S_i)$ for each bootstrap sample S_i.

3 Use the bootstrap statistics $\{D(S_i)\}_{i=1}^B$ to estimate $P(D \geq D(S))$, and accept the null hypothesis at a significant level α if $P(D \geq D(S)) \geq \alpha$.

The multi-dimensional density estimation in general is not as reliable as univariate regression, hence we do not expect that the density estimation method would outperform the regression method. Nevertheless, as a comparison to the regression approach, we would like to conduct a simulation study of the density estimation approach.

3. Simulation

3.1 Purpose of simulation study

Given the two approaches to the nonparametric conditional independence testing, we would like to ask the following questions:

1. Are these two approaches to nonparametric conditional independence testing practical?
2. Which of the two proposed approaches has a better performance?
3. If an algorithm is not always reliable, can we identify the cases where it is reliable?

The best way to answer these questions is to conduct a simulation study. This is because, first, with synthetic data, we know whether a certain conditional independence relation holds among a set of variable, while for real data, we usually do not know. Second, it is easier to generate a relatively large number of synthetic data sets than to collect the same number of real data sets.

3.2 Implementation

3.2.1 Data generation

We generate 99 data sets, each of sample size 250, from 99 additive models. For the density estimation

[10] The reason why we do not directly use all the information to test the conditional independence relation is that the BIC score is well defined, and has desirable statistical interpretations.

approach, we also generate 45 samples, each of sample size 2000, from 45 additive models. Each sample has 5 variables. The functional relations among the variables are linear combinations of exponential, log, trigonometric, rational, and polynomial functions.

3.2.2 Implementation of the regression approach

The algorithm for the regression approach to the nonparametric conditional independence testing is implemented in Splus. The Splus package has a built-in function `gam`, which is an implementation of the back-fitting additive regression algorithm proposed in Hastie & Tibshirani (1990). It also has the univariate smoothing function `smooth.spline`, which will be used as the univariate regression module for the `gam` function (Venables & Ripley 1997).

The reliability testing is also implemented in Splus. The Splus function `gam`, with the link function set to `logit`, is used to learn the models for reliability prediction.

3.2.3 Implementation of the density estimation approach

The algorithm for the density estimation approach is implemented in C and Splus. The Splus codes provide an interface for data importing and exporting. The computation is done by the C routines called by the Splus interface.

To estimate the marginal and joint densities, we use the one-, two-, and three-dimensional Epanechnikov kernels (Haerdle, 1990). Each variable is normalized so that it has mean 0 and standard deviation 1. The densities are estimated at each data point. To compute the test statistic, we use the following formula:

$$\hat{D} = \sum_{i=1}^{N} \left(1 - \sqrt{\frac{\hat{f}_{x,z}(x_i, z_i)\hat{f}_{y,z}(y_i, z_i)}{\hat{f}_z(z_i)\hat{f}_{x,y,z}(x_i, y_i, z_i)}} \right)^2$$

where N is the sample size, (x_i, y_i, z_i) are the values of the (X, Y, Z) for the i^{th} data point, and $\hat{f}_{x,z}(x, z), \hat{f}_{y,z}(y, z), \hat{f}_z(z)$, and $\hat{f}_{x,y,z}(x, y, z)$ are estimated densities for (X, Z), (Y, Z), Z, and (X, Y, Z) respectively.

The bootstrap samples are generated in the following way:

1. For each z_i, generate an x'_i such that:
$$P(x'_i = x_j) = \frac{f_{x,z}(x_j, z_i)}{\sum_{k=1}^{N} f_{x,z}(x_k, z_i)}.$$
2. Independently, for each z_i, generate a y'_i such that:
$$P(y'_i = y_j) = \frac{f_{y,z}(y_j, z_i)}{\sum_{k=1}^{N} f_{y,z}(y_k, z_i)}.$$
3. $\{(x'_i, y'_i, z_i)\}_{i=1}^{N}$ will be a bootstrap sample generated according to the null hypothesis $X \perp\!\!\!\perp Y | Z$.

3.3 Empirical results

3.3.1 Regression approach

Using the algorithm for the regression method, we conducted 347 tests of conditional independence. Among them, there are 148 tests where the null hypothesis should be accepted, i.e., the conditional independence relation holds. For the other 199 tests, the null hypothesis should be the rejected, i.e., the conditional independence relation does not hold. The actual testing results are given in table 1 and table 2.

Table 1. Regression Approach: Null hypothesis is true

number of tests	rate of type I error
148	24.3%

Table 2. Regression Approach: Null hypothesis is false

number of tests	rate of type II error
199	46.7%

These two tables show that the general accuracy of the regression approach is not satisfactory, hence this approach would be useless if we could not tell if a test is reliable.

To learn to predict the reliability of the conditional independence tests, we choose 248 tests as training data. The remaining 99 tests are used as testing data. Following are the results of the reliability test:

1. Among the 99 tests, 57 of them accepted the null hypothesis:
   - 21 tests are predicted reliable, and they are all correct.
   - 36 tests are predicted unreliable, 15 of them are incorrect.
2. Among the 99 tests, 42 of them rejected the null hypothesis:
   - 34 tests are predicted reliable, and they are all correct.
   - 8 tests are predicted unreliable, 7 of them are incorrect.

The above results suggest that, first, if a test is predicted reliable, then highly likely it is correct; and second, a significant number of tests are predicted reliable. Therefore, combining the nonparametric conditional independence testing algorithm with the reliability prediction procedure, in many cases we can provide very accurate conditional independence information.

3.3.2 Density estimation approach

Compared to the results of the regression approach, the results of the density estimation approach are disappointing. We conducted 200 conditional independence tests on 50 samples of size 250. The results for these test were bad. So we generated 45 samples of size 2000, and conducted another 90 tests on these larger samples. The testing results are given in table 3 and table 4.

Table 3. Density Estimation: Null hypothesis is true

number of tests	sample size	type I error
100	250	15%
45	2000	40%

Table 4. Density Estimation: Null hypothesis is false

number of tests	sample size	type II error
100	250	60%
45	2000	26.7%

From these results we can see that, first, the density estimation approach performed badly, and second, increasing the sample size does not help. The poor performance of the density estimation approach may be partly due to the fact that the density estimation is an unsupervised learning, and partly to the fact that the bootstrap sampling method, which is used in our density estimation approach algorithm, is not very reliable for this type of application. We doubt there is any way we can predict the reliability of the density estimation approach as we did to the regression approach.

4. Prior Work

4.1 Conditional Independence Test for Discrete Data

Conditional independence test for a set of discrete random variables has been well studied.[11] Consider three discrete random variables: $V = \{X, Y, Z\}$, suppose X has m levels, Y n levels, and Z p levels. To test the null hypothesis that $X \perp\!\!\!\perp Y | Z$, we can use the following statistic:

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} \frac{\left(N_{X_i,Y_j,Z_k} - \hat{N}_{X_i,Y_j,Z_k}\right)^2}{\hat{N}_{X_i,Y_j,Z_k}} \quad (1)$$

where N_{X_i,Y_j,Z_k} is the number of observations with $X = X_i$, $Y = Y_j$, $Z = Z_k$; N_{X_i,Z_k} the number of observations with $X = X_i$ and $Z = Z_k$; N_{Y_j,Z_k} the number of observations with $Y = Y_j$ and $Z = Z_k$, N_{Z_k} the number of observations with $Z = Z_k$, and $\hat{N}_{X_i,Y_j,Z_k} = \frac{N_{X_i,Z_k} N_{Y_j,Z_k}}{N_{Z_k}}$. For large samples, Q has approximately a χ^2 distribution with degrees of freedom: $(m-1)(n-1)p$.

4.2 Conditional Independence Test for Linear Model

For linear models with a normal distribution, a function of the sample partial correlation could be used as a statistic for conditional independence test.[12] For random variables X, Y, and Z, the sample partial correlation of X and Y, with Z held constant, is:

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}} \quad (2)$$

where r_{xy}, r_{xz}, and r_{yz} are sample correlations for X and Y, X and Z, and Y and Z respectively. Under the null hypothesis, i.e., the partial correlation $\rho_{xy.z} = 0$, the follow statistic has a t distribution with $n-3$ degrees of freedom:

$$t = \frac{r_{xy.z}\sqrt{n-3}}{\sqrt{1 - r_{xy.z}^2}} \quad (3)$$

where n is the sample size.

Alternatively, we could use Fisher's z statistic:

$$z(\rho_{xy.z}, n) = \frac{1}{2}\sqrt{n-4}\log\frac{1+\rho_{xy.z}}{1-\rho xy.z} \quad (4)$$

where $\rho_{xy.z}$ is the population partial correlation. The distribution of $z(\rho_{xy.z}, n) - z(r_{xy.z}, n)$ has a standard normal distribution.

[11] Chapter 4, Conover (1980).

[12] §3.4, Jobson (1991).

4.3 Nonparametric Conditional Independence Test

However, whenever the data sets are continuous but not necessarily generated from linear models, we have to apply nonparametric testing method to obtain conditional independence information. Unfortunately, there is not much literature about nonparametric conditional independence test.[13] The only paper on this topic I have ever found is Linton and Gozalo (1999). However, the algorithm 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 I error. That is, it is highly likely that the test will reject the null hypothesis (conditional independence) even if it is true.

4.3.1 Extension of Current Methods

One approach to the nonparametric conditional independence testing is trying to extend the available nonparametric independence tests between two random variables, such as the Spearman's ρ test and Kendall's τ test[14], Kallenberg and Ledwina (1999), Gieser and Randles (1999), and Kano (1992). However, these algorithms cannot be extended to test conditional independence relations, unless the conditioning variable is discrete. The reason is simple: the fact that X and Y are independent conditional on $Z = z$ for every z in the codomain of Z does not imply that, for any measurable subset A of the codomain of Z, X and Y are conditionally independent given $Z \in A$.[15] Therefore, 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.

Another approach is to try to extend the available conditional independence testing algorithms for discrete variable or linear model. People have routinely converted the continuous data into discrete data when the algorithms they need are available only for the discrete data (Friedman & Goldszmidt, 1996). The problem with this approach, again, 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. 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.

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 conditional independence information. However, the current methods for conditional independence test can only be applied to purely discrete data or continuous data generated from linear models.

The lack of nonparametric methods for conditional independence testing has largely confined the application of causal inference to a restricted set of data sets. In this paper we proposed two nonparametric conditional independence testing algorithms that could be applied to continuous data with unknown functional relations: the regression approach and the density estimation approach. The simulation study of these two approaches showed that, while not suitable for all cases where nonparametric tests are needed, the regression approach, combined with a reliability testing algorithm, could give very good results for a large and identifiable sets of data. The density estimation approach, while theoretically suitable for a broader set of data, performed badly.

The simulation study showed that the regression approach has a great potential as a nonparametric conditional independence testing method. There is still plenty room of improvement for the algorithm proposed in this paper. For example, we could try other regression methods, such as surface smoother, wavelet, etc, in the algorithm for conditional independence testing. We could also use other supervised learning methods to learn to predict the reliability of the conditional independence tests. Finally, a larger collection of training data will definite improve the performance of the reliability testing algorithm.

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

[14] Chapter 5, Conover (1980).

[15] 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!

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.

Biddle, J., & Zarkin, G. (1989) "Choice among Wage-Hours Packages: An Empirical Investigation of Male Labor Supply," *Journal of Labor Economics*, 7, pp.415-437.

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.*

Haerdle, W. (1990) *Smoothing Techniques with Implementation in S*, New York: Springer-Verlag.

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., & 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.*

Venables, W., & Ripley, B. (1997) *Modern Applied Statitsics with S-PLUS*, 2nd ed., New York: Springer.

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.