Simulation Study of Conditional Independence Test Using GAM Method

Chu, Tianjiao

December 13, 2000

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 prevent 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. Works by Pearl, Spirtes, Glymour, Scheines, and other people,[1] has shown that we could gain some insight of the causal relations among random variables with the knowledge of conditional independence relations among these variables.

In this paper, I will report a simulation study of extracting conditional independence information from observational data using a regression method, the generalized additive model (GAM). I will first briefly a current approach to the inference of causal relations based on conditional independence information, as well as the current method of testing conditional independence relation. Then I will discuss the GAM approach to the conditional independence test for continuous data. I will report the process the generating simulation data, and the conditional independence test result of applying the GAM method to these data. I will also consider how well as the how well these tests could be used to infer causal information. I will also give an analysis of the performance of the GAM method. Finally, I will discuss possible future work.

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

2 Causal Graph and Conditional Independence Test

2.1 Causal Graph

A directed graph is defined as a set of vertices and a set of directed edges connecting pairs of vertices. If we require that these edges cannot form any directed cycle, i.e., we cannot start from a vertex X, follow the direction of the edges, and end at X, then the graph is called a directed acyclic graph (DAG). DAGs provide an intuitive representation of causal relations among a set of random variables: Each vertex in the graph represents a random variable, and a variable X is a direct cause of a variable Y if and only if there is a direct edge from X to Y in the graph.

We will further impose the following two assumption to a directed acyclic graph:[2]

Markov Condition: If the causal relations among a set of random variables for a population can be represented by a graph G, then a variable X is independent of a set of variables $\boldsymbol{Y}$ conditional on another set of variables $\boldsymbol{Z}$, provided that, in the corresponding graph G, the vertices representing variables in $\boldsymbol{Y}$ are neither parents, nor descendents of the vertex representing X in G, and the vertices representing variables in $\boldsymbol{Z}$ are the set of all the parents of the vertex representing X in G.

Faithfulness Condition: If we observe that, in a population, a random variable X is independent of a random variable Y given a set of random variables $\boldsymbol{Z}$, this conditional independence relation follows from the corresponding causal graph G by the Markov Condition. For example, if X and Y are conditional independent given some other set of variables, the vertices representing X and Y respectively in G cannot be connected by an edge.

These two conditions establish a close relation between causation and conditional independence. Based on this relation, we can infer causal information from knowledge about conditional independence. The SGS algorithm, for example, can output the set of all of the possible causal graphs that are compatible with the conditional independence relations observed in a population, provided that there is no unobserved variable that is the parent of any two observed variables in the population.

Further studies have been focused on causal inference with the presence of unobserved variables that are parents of pairs of observed variables. Algorithms, such as FCI, have been developed to infer common causal patterns from populations sharing same set of conditional independence relations among observed variables. Moreover, causal inference from the population where variables may be causes of each other have also been studied.[3]

[2] Spirtes et al (2000).
[3] Richardson (1996).

2.2 Current Tests for Conditional Independence

If we take a look at the kind of data where people try to draw causal information by the method mentioned above, we will find that these data are either discrete (or have been discretized before use), or assumed to be linear in terms of the functional relation between a child and its parents. This is not surprising, because basically the available methods of conditional independence test can only be applied to either discrete data, or data where the variables are linearly connected.[4]

2.2.1 Conditional Independence Test for Discrete Data

Conditional independence test for a set of discrete random variables has been well studied.[5] 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:

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

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 $\hat{N}_{X_i,Y_j,Z_k} = \dfrac{N_{X_i,Z_k} N_{Y_j,Z_k}}{N_{Z_k}}$. For large samples, Q has approximately a χ^2 distribution with degrees of freedom: $(m-1)(n-1)p$.

If we are not satisfied with the χ^2 approximation for small sample sizes, we can apply Monte Carlo method to simulate the distribution of the statistic Q under the null hypothesis.

2.2.2 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. [6]For random variables X, Y, and Z, the sample partial correlation of X and Y, with Z held constant, is:

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

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:

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

[4] Loglinear Case
[5] Chapter 4, Conover (1980).
[6] §3.4, Jobson (1991).

where n is sample size.

Alternatively, we could use Fisher's z statistic:

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

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.

2.3 Nonparametric Conditional Independence Test

However, we often need to know the conditional independence relations among a set of continuous variables, or a mixture of continuous and discrete variables, where we do not know very much about the functional relations among the variables, nor do we know much about the distribution of each error term. In this case, non-parametric conditional independence tests would be applied.

2.3.1 Nonparametric Independence Test

The nonparametric Independence test is a special case of nonparametric conditional independence test, where the set of variables to be conditioned on is empty. The fact that we only need to consider two random variables makes the test much simpler than the conditional independence test in general.

When X and Y are both continuous, we can use either Spearman's ρ test or Kendall's τ test to find whether X is independent of Y.

Here we use H_0 to denote the hypothesis that X and Y are independent. Let n be the sample size. If we use the ρ test, the Spearman's statistic is defined as:

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\left(\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{\frac{1}{2}}} \tag{5}$$

Let N_c be the number of pairs (X_i, Y_i) and (X_j, Y_j) where $\frac{Y_j - Y_i}{X_j - X_i} > 0$ plus half of the number of pairs where $Y_j = Y_i$, and N_d the number of pairs where $\frac{Y_j - Y_i}{X_j - X_i} < 0$ plus half of the number of pairs where $Y_j = Y_i$. The Kendall's statistic is defined as:

$$\tau = \frac{N_c - N_d}{N_c + N_d} \tag{6}$$

A level α test will reject the null hypothesis if the value of the ρ or τ statistic exceed their $1 - \frac{\alpha}{2}$ quantiles or are less than their $\frac{\alpha}{2}$ quantiles respectively.

2.3.2 Extension of Current Conditional Independence Tests

One approach to non-parametric conditional independence testing is trying to extend the available non-parametric independence test between two random variables. Unfortunately, these tests cannot be extended to conditional independence tests, 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 $Z \in A$.[7] 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 test for discrete variable or linear model. People have routinely converted the continuous data into discrete data, and then applied conditional independence tests for discrete data. The problem with this approach is that the conditional independence relation among the continuous variables does not imply conditional independence among the discretized variables.[8] Conversely, conditional independence among discretized variables does not imply conditional independence among continuous variables either.[9] 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.

2.3.3 Conditional Independence Test by Density Estimation

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 discretization method discussed above is also a kind of density estimation. However, in this section, we are concerned with continuous density estimation.

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.

[7]For example, suppose $Z \sim U(0,100)$, $X \sim U(Z-1, Z+1)$, $Y \sim U(Z-1, Z+1)$, and $X \perp\!\!\!\perp Y | Z = z$. It is easy to see, however, that $X \not\perp\!\!\!\perp Y | Z \in (0,50)$. Without knowledge of Y, we only know that X lies in the interval $(-1, 51)$, an interval of length 52. With the knowledge of Y, we know that X lies in $(\max[Y-2,-1], \min[51, Y+2])$, an interval of length at most 4.

[8]Use the example in the previous footnote, it is easy to see that:
$Pr(X \in (0,1), Y \in (10,11) | Z \in (0,50)) = 0 < Pr(X \in (0,1)|Z \in (0,50)) Pr(Y \in (10,11)|Z \in (0,50)$

[9]Suppose $X \sim U(-1,1)$, $Y \sim U(-1,1)$, and $X \perp\!\!\!\perp Y$. Suppose $Z = X + Y - 20$ if $X < 0, Y < 0$; $Z = X + Y - 10$ if $X < 0$, $Y \geq 0$; $Z = X + Y$ if $X \geq 0$, $Y < 0$; and $Z = X + Y + 10$ if $X \geq 0$, $Y \geq 0$. It is easy to see that $X \not\perp\!\!\!\perp Y | Z = z$. For example, given $Z = 0$, we know that $X \in [0,1)$; but given $Z = 0$ and $Y = -0.5$, we know that $X = 0.5$. However, conditional on $Z \in (k-2, k+2)$, where $k = -20, -10, 0, 10$, X and Y are independent. Hence $Pr(X \in A, Y \in B | Z \in (k-2,k+2)) = Pr(X \in A|Z \in (-2,2)) Pr(Y \in B|Z \in (-2,2))$, where A, B are measurable subsets of the real line.

The distance D between two densities $f_{x,y,z}(x,y,z)$ and $\dfrac{f_{x,y}(x,y)f_{y,z}(y,z)}{f_z(z)}$ can be any distance measure. Let $D = \int \delta(x,y,z)\,dx\,dy\,dz$, then one choice of δ is:

$$\delta(x,y,z) = \left(\sqrt{\hat{f}_{x,y,z}(x,y,z)} - \sqrt{\frac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}}\right)^2 \tag{7}$$

which is the kernel of Hellinger distance.

We may also want to use

$$\delta(x,y,z) = \left|\hat{f}_{x,y,z}(x,y,z) - \frac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}\right| \tag{8}$$

which is the kernel of L_1 distance, or

$$\delta(x,y,z) = \left(\hat{f}_{x,y,z}(x,y,z) - \frac{\hat{f}_{x,y}(x,y)\hat{f}_{y,z}(y,z)}{\hat{f}_z(z)}\right)^2 \tag{9}$$

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 data $S_1, \cdots, S_N$ 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 $F_d(D) \leq \alpha$.

Despite its conceptual appeal, the density estimation method 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.

3 Conditional Independence Test with GAM

In this paper, we will use the method of Generalized Additive Model (GAM), which belongs to the family of regression methods, to test conditional independence. As we are going to see, this method provides an approach that has much weaker assumptions than the linear normal model, and is not subject to the curse of dimensionality as with the density estimation method.

3.1 Regression Approach to Conditional Independence Testing

The basic idea of the regression approach to conditional independence testing is that, if a variable Y is conditionally independent of X given a set of variables Z, then the knowledge of Y does not help us predict X if we already know Z. In practice, if we want to test whether X is independent of Y given Z, we would assume that X is not a constant, and is a smooth function of Y and Z, plus a certain error term ϵ. Usually, due to the lack of information about the error term ϵ, we further assume that ϵ is $N(0, \sigma^2)$, with σ^2 unknown.

The general algorithm of regression method is:

1 Regress X on Z and $\{Y, Z\}$ respectively.

2 Score the fits of X on Z and on $\{Y, Z\}$ using some kind of score function.

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

The rationale behind the regression approach is that if two variables X and Y are independent, then we are not going to predict X better even if we know Y. That is, the variance of X is the same as the expected value of the variance of X given Y, if X and Y are independent.

In general the variance of X is greater than or equal to the expected variance of X given Y, that is:

$$\mathrm{E}[\mathrm{Var}(X|Y)] \leq \mathrm{Var}[X] \tag{10}$$

To show this, we notice that:

$$\begin{aligned}\mathrm{E}[\mathrm{Var}(X|Y)] = \mathrm{E}[\mathrm{E}[X^2|Y] - (\mathrm{E}[X|Y])^2] = \mathrm{E}[\mathrm{E}[X^2|Y]] - \mathrm{E}[(\mathrm{E}[X|Y])^2] = \mathrm{E}[X^2] - \mathrm{E}[(\mathrm{E}[X|Y])^2] \\ \leq \mathrm{E}[X^2] - (\mathrm{E}[\mathrm{E}[X|Y]])^2 = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{Var}[X]\end{aligned}$$

for $\mathrm{E}[(\mathrm{E}[X|Y])^2] = \mathrm{Var}(\mathrm{E}[X|Y]) + (\mathrm{E}[\mathrm{E}[X|Y]])^2 \geq (\mathrm{E}[\mathrm{E}[X|Y]])^2$.

Note that although independence implies that $\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|Y)]$, the converse does not hold. However, it is not difficult to show that if the variance of X is equal to the expected variance of X given Y, then the expected value of X is independent of Y:

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

$$\implies \mathrm{Var}(\mathrm{E}[X|Y]) = 0 \implies \mathrm{E}[X|Y] = \mathrm{E}[\mathrm{E}[X|Y]] = \mathrm{E}[X]$$

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 $\mathrm{E}[X] = \mathrm{E}[X|Y] = 0$.

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 is the parent of X, then $\mathrm{E}[X|Y] = f(Y)$, 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, change the order of the regression, i.e., regress Y on X, will solve the problem.

The above argument can be applied to the case of conditional independence. Let $\boldsymbol{Z}$ be a set of random variables, $\mathrm{E}[\mathrm{Var}(X|Y, \boldsymbol{Z})] = \mathrm{Var}(X|\boldsymbol{Z})$ if $X \perp\!\!\!\perp Y|\boldsymbol{Z}$. Again, $\mathrm{E}[\mathrm{Var}(X|Y, \boldsymbol{Z})] = \mathrm{Var}(X|\boldsymbol{Z})$ does not imply $X \not\perp\!\!\!\perp Y|\boldsymbol{Z}$, but does imply $\mathrm{E}[X|\boldsymbol{Z}] = \mathrm{E}[X|Y, \boldsymbol{Z}]$.

3.2 Choice of Regression Methods

1. Surface Smoother

   Among the regression methods, the surface smoother method imposes the least restriction on the relations between variables. Here we only assume that:

   $X = f(Y, Z) + \epsilon$,

   where f is an arbitrary smooth function, and ϵ is independent of (Y, Z).

   This method works for the case where we only have 2 or 3 predictors in the model. However, as the number of predictors increases, this method is subject to the curse of dimensionality.

2. 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. In our 3 variable case, this means:

   $X = f(Y) + g(Z) + \epsilon$,

   where f and g are arbitrary smooth functions, and ϵ is independent of (Y, Z).

   This method avoids the problem of dimensionality, because it regresses the response variable X on Y and Z separately. Later we will discuss this approach with more details.

3. Other Methods

If we want to relax the restrictions imposed by the additive model, and to some extent retain its performance for high dimension cases, we may end up with some kind of mixture of the additive model and surface smoother.

For example, let $\vec{Y}$ be a vector of predictors, the Projection Pursuit method assumes that:

$$X = \sum_{j=1}^{p} g_j(\beta_j^T \vec{Y}) + \epsilon,$$

where $\vec{Y}$ is a vector of predictors, g_j's are arbitrary smooth functions, and ϵ is independent of (Y, Z).

The advantage of Projection Pursuit is that it is almost as general as Smoother Surface. However, it is computationally expensive.

We may also want to use low dimension smoother surface regression in an additive model. For example, if we have 4 variables X, Y, Z, and U, we may assume that:

$$X = f(Y, Z) + g(Z, U) + h(U, Y) + \epsilon.$$

where f, g, h are arbitrary smooth functions, and ϵ is independent of (Y, Z).

3.3 Generalized Additive Model Regression

The generalized Additive Model (GAM) is slightly more general than the additive model we mentioned above. The GAM assumes that a univariate function, called the link function, of the response variable, is a linear combination of smooth (univariate) functions of each predictor. (In fact, GAM allow function of more than one predictor, but in this paper, we will only consider the univariate function.) In this paper, we will suppose that the link function is the identity function, unless the response variable is nominal discrete, in which case the link function is assumed to be the `logit` function, i.e., $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$.

Suppose we want to regress a response variable Y on a set of predictors $X_1, \cdots, X_n$, where $Y, X_1, \cdots, X_n$ are all continuous, by GAM method. This means that we assume that $\text{E}[Y|\boldsymbol{X}] = \sum_{i=1}^{k} f_i(X_i)$. We would choose a smoother, such as cubic spline smoother,[10] for each of the predictors X_i. This smoother is used to estimate the smooth function f_i. Presumably, we should also set a smoothing parameter for each smoother. (The smoothing parameter will determine how to balance the bias and variance of smoothing.) Then we begin an iterative algorithm, called backfitting, as described below:

[10] A cubic spline smoother will return a piecewise cubic polynomial function that is twice continuously differentiable. This function is the solution to a variational problem that minimizes the sum of two terms:

$$S_\lambda(g) = \sum_{i=1}^{n}(Y_i - g(X_i))^2 + \lambda \int (g''(x))^2 dx$$

where Y_i and X_i are values of the i^{th} observation of the response and predictor variables respectively, and λ is smoothing parameter. The g that minimizes S_λ is the result of cubic spline smoothing.

Let $\hat{f}_i^m$ denote the estimation of f_i at the m^{th} iteration,

- Initialize $\hat{f}_i^0$ for $i = 1, \cdots, k$. Let $m = 1$
- For $j = 1, \cdots, k$, smooth $Y - \left(\sum_{i=1}^{j-1} \hat{f}_i^m(X_i) + \sum_{i=j+1}^{k} \hat{f}_i^{m-1}(X_i) \right)$ against X_j, and let the $\hat{f}_j^m$ be the newly computed smoothing curve.
- If $\hat{f}_i^m = \hat{f}_i^{m-1}$ for all i, stop. Otherwise, increase m by 1.

3.4 Score Functions for GAM Model

AIC and BIC scores are two of the most popular score functions used in model selection.[11] AIC score is designed for the search of the model that minimizes prediction error, while BIC score is designed for the search of the model with highest posterior. When used to compare alternative multiple linear regression models, the AIC and BIC are defined as:

$$AIC = RSS + 2\,df\,\sigma^2 \tag{11}$$

$$BIC = RSS\,/\,\sigma^2 + \log(N)\,df \tag{12}$$

where RSS is the residual sum of square, df the degree of freedom, i.e., the number of parameters in the regression model, N the sample size, and σ^2 the noise variance.

The formula for the noise variance σ^2 is:

$$\sigma^2 = RSS_0\,/\,(N - df_0) \tag{13}$$

where RSS_0 and df_0 are obtained from the most complicated model for the response variable, i.e., from the model that includes all the random variables in the sample, except the response variable, as predictors.

In the multiple linear regression case, asymptotically, AIC will lead to the model with smallest prediction error, and BIC will lead to the true model. By replacing the degrees of freedom with the equivalent degrees of freedom, we can extend AIC and BIC to the non-parametric regression, which is the core of the GAM regression. Because the equivalent degree of freedom does not reflect the number of free parameters in the model, the asymptotic properties of AIC and BIC are lost in the GAM regression model. For linear smoothers, which include the spline smoother used in this study, one way to define the equivalent degree of freedom is to use the trace of the smoother matrix. (A smoother is linear if when we use it to smooth a linear combination of two response variables Y_1, Y_2 against a set of predictors $\boldsymbol{X}$, the resulting smooth curve

[11] Akaike (1973), Schwartz (1978).

is the linear combination of the two smoothing curves obtained by smooth Y_1 and Y_2 against $\boldsymbol{X}$ separately). The smoother matrix $\boldsymbol{S}$ for a linear smoother is an $n \times n$ matrix such that:

$$\hat{Y} = \boldsymbol{S}Y \tag{14}$$

where Y is the response variable, $\hat{Y}$ is the vector of fitted values of the response variable, and n is sample size.

AIC and BIC scores are not the only ways to do model selection.[12] Another popular choice is the χ^2 test. The basic idea of the χ^2 test is that, assuming a certain set of regularity conditions being satisfied, the deviance between two nested alternative models has asymptotically χ^2 distribution with degrees of freedom equal to the difference of free parameters in the two models. However, we are not going to use χ^2 test in this simulation study. The main reason is that, in our simulation setting, the difference of equivalent degrees of freedom between two alternative models is usually fixed. Therefore, at any given level, the χ^2 test is equivalent to test whether the difference of deviances of the two models is greater than a constant value. As we are going to see in later sections, this is too simple to be a test of conditional independence. Another reason why we do not use χ^2 test is that the distribution theory of deviance of the GAM models is not developed.[13]

3.5 GAM Method for Conditional Independence Testing

Consider a DAG G with a set of vertices $\boldsymbol{V}$. By a GAM interpretation of G, or a GAM model based on G, is meant a set of random variables corresponding to the vertices $\boldsymbol{V}$ and a set of restrictions imposed on these random variables: Each child variable, i.e, a variable corresponding to a child vertex in G, is equal to a linear combination of smooth univariate functions of their parents, plus an independent error term.

When using GAM method to test conditional independence, we usually assume the sample data were generated from a GAM model based on a certain DAG. Suppose the sample contains the following random variables: $X, Y, Z_1, \cdots, Z_k, U_1, \cdots, U_m$. If we want to test whether X is independent of Y given $(Z_1, \cdots, Z_k)$, ($k \geq 0$, if $k = 0$, this will be a test of whether X and Y are independent), we can apply GAM regression to the sample data according to the following two alternative regression models:

$$M_{Y,Z}: \quad X = f(Y) + \sum_{i=1}^{k} g_i(Z_i) + \epsilon \tag{15}$$

$$M_Z: \quad X = \sum_{i=1}^{k} h_i(Z_i) + \epsilon \tag{16}$$

[12] George (2000).
[13] p.155, Hastie et al. (1990).

We then compute the AIC and BIC scores for the two models respectively. Let $AIC_{Y,Z}, BIC_{Y,Z}$ be the scores for model $M_{Y,Z}$, and AIC_Z, BIC_Z be the scores for model M_Z. Then, if we use the AIC score, we would claim that X and Y are independent conditional on $\boldsymbol{Z}$ if and only if $AIC_Z \leq AIC_{Y,Z}$. Similarly, if we use BIC score, we would claim that X and Y are independent conditional on $\boldsymbol{Z}$ if and only if $BIC_Z \leq BIC_{Y,Z}$.

Note that BIC and AIC criteria differ only at how they penalize the model complexity. More precisely, BIC penalizes the model complexity by multiplying the degrees of freedom by $\log(n)$, where as AIC multiplies the degrees of freedom by 2. Therefore, in case the results given by AIC and BIC scores differ, the BIC score always prefers the simpler regression model, and the AIC score always prefers the more complex one.

3.6 Comments on GAM Method

One big concern with GAM is the assumption that the functional relation between the response variable and its predictors is additive may be not always reasonable. If this assumption does not hold, we will have the following problems:

Suppose we want to test whether $X \perp\!\!\!\perp Y \mid (Z_1, \cdots, Z_k)$. If we choose $X \not\perp\!\!\!\perp Y \mid (Z_1, \cdots, Z_k)$ as the null hypothesis, the violation of the additive assumption means that, under the null hypothesis, X is not a linear combination of smooth functions of Y and $Z_1, \cdots, Z_n$. Therefore, the GAM test may not be reliable.

Now suppose that we let $X \perp\!\!\!\perp Y | (Z_1, \cdots, Z_k)$ be the null hypothesis. This is fine as long as $k = 1$ contains no more than one random variable. If $(Z_1, \cdots, Z_k)$ is a vector of 2 or more random variables, when the additive assumption is violated, even when the null hypothesis is true, the GAM test will not be reliable.

Unfortunately, the violation of the additive assumption is almost inevitable when we use the GAM method to derive the conditional independence relations among a set of random variables from some sample data, even if the sample data were generated from a GAM model based on a DAG, i.e., each child variable is a linear combination of smooth univariate functions of its parents plus some independent error term.

To illustrate this point, consider the DAG G in figure 1. Suppose $X_4 = f_1(X_1) + f_2(X_2) + f_3(X_3) + \epsilon_1$, $X_5 = g(X_4) + \epsilon_2$, and $X_1, X_2, X_3, \epsilon_1$, and ϵ_2 are independent normally distributed random variables with mean 0. If we only want to test whether $X_4 \perp\!\!\!\perp X_1 | (X_2, X_3)$, the additive assumption is not violated. However, we also need to test whether $X_5 \perp\!\!\!\perp X_1 | (X_2, X_3)$, etc. It is easy to see that, in general, the additive assumption does not hold for the latter test, because, in general, X_5 is not equal to a linear combination of smooth functions of X_1, X_2, and X_3, plus some independent error term. (Consider the case where f_1, f_2, f_3 are the identity function, and g the square function.)

Further more, the marginalization of a GAM model in general is no longer a GAM model. That is, if we allow the existence of latent variables, we may find that the set of observed variables cannot be interpreted

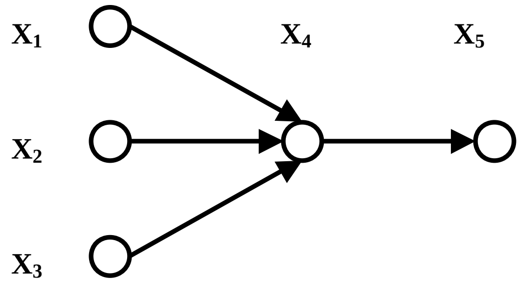

Figure 1: G

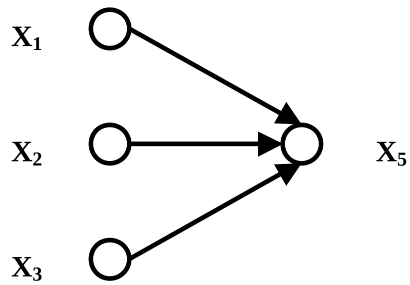

Figure 2: G'

as a GAM model based on some DAG. For example, consider the GAM model mentioned in the previous paragraph. Suppose now that X_4 is a latent variable, then conditional independence relations among the 4 observed variables X_1, X_2, X_3 and X_5 is encoded by the DAG G' in figure 2. However, it is easy to see that, in general, no GAM interpretation of G' has the joint distribution as the marginal distribution of (X_1, X_2, X_3, X_5).

Despite these limitations, GAM method is still a promising approach, because, first, the additive assumption, although not always reasonable, is much weaker than than that of the linear normal model, hence general enough for many situations. Second, with the additive assumption, we do not need to worry about the curse of dimensionality. Third, as we are going to see, that the violation of the GAM assumptions does not necessarily lead to undesirable test result.

4 Simulation

To do the simulation study, we built a GAM data generator, and used this program to generate more than one thousand samples based on some GAM models. We also wrote S codes for conditional independence tests, and implemented the PC algorithm in S. We then imported the data to the S-plus program and conducted conditional independence tests, as well as causal pattern search. Finally the results were output and analyzed.

4.1 GAM Model Specification

In this paper, by a GAM model we do not mean a regression model such as $Y = f(X) + \epsilon$, although the latter can be viewed as a special kind of GAM model. A GAM model is a system that includes a set of random variables, and the information about which of them are continuous, or discrete, and for discrete variables, their number of levels. A GAM model also contains a set of causal relations connecting the random variables such that a DAG representing these relations does not contain a directed cycle, as well as a set of functional relations between each random variable X and the set of random variables in the model that are direct causes of X. Finally, for each random variable that has no direct cause in the model, the model must specify its distribution, and for each continuous variable that has direct causes in the model, the model must specify the distribution of an additive error term attached to that variable.

Here we need to given a brief explanation of how each sample point of the continuous and discrete data is generated so that the following description of the structure of GAM data generator makes sense.

- A continuous variable with no direct cause in the model is generated from a given distribution by a random number generator. So are the additive terms to a continuous or ordinal discrete variable.

- A continuous variable with direct causes in the model is generated by computing a function of its direct causes, and then adding an error term to the result. Note that if a direct cause is a nominal discrete variable, its values will be converted to integers before being placed in the function.

- A discrete variable with no direct cause in the model is generated from a multinomial distribution by a random generator.

- A discrete variable with direct causes in the model is generated is generated in the following way:

  First, for each level of the variable, compute a distinct function of the direct causes of the variable. Then the values of all the levels are converted to a multinomial probability distribution over the all the levels. Finally, the value of this variable is generated from that distribution by a random generator.

4.2 GAM Data Generator

An essential part of this simulation study is to build a GAM data generator. The basic requirement for a GAM data generator is that it must be both able to generate any number of samples automatically according to some pre-specified model, and able to generate random GAM models automatically, then generate samples from these models.

The GAM data generator contains three parts: a graph reader, a model generator, and a sample generator.

Graph Reader: The function of graph reader is to parse a file that describes the vertices and edges of a DAG, and create a DAG according to that file, where a file specifying a DAG containing lines that specify, for each vertex in the DAG, the set of vertices, if any, that are parents of that vertex in the DAG. Of course, it should check whether the file is a valid description of a DAG. For example, it will check whether there is any directed cycle implied in the description.

Model Generator: The heart of the GAM data generator is the model generator, which either generates a GAM model based on a file that completely specifies that model, or generates a semi-random GAM model based on a file partially specifying a GAM model, or generates a random GAM model based on a file specifying a DAG.

To specify a GAM model completely, besides specifying the DAG, we need also the following information for each variable in the model:

- We need to specify the type of variable, which could be continuous or discrete.
- If a variable is discrete, we need to specify the number of levels. The lower bound is 2, and the upper bound is 12.
- If a variable has no direct cause in the model, we need to specify the distribution of the variable.

  If the variable is continuous, it may take any of the 4 families of distribution: Uniform, Normal, Gamma, and Mixture of two normal with the same variance and same weight. The mean and variance of the distribution, given the knowledge of the distribution family, will be used to determine the distribution uniquely.

  If the variable is discrete, we need to specify the probability for the variable in each level.
- If a variable is continuous, and has direct causes in the model, we need to specify the distribution of the additive error term. Here again we allow 4 families of distributions: Uniform, Normal, Normalized Gamma (a Gamma random variable minus its mean), and Mixture of two normal with same variance and same weight. The mean of the error term must be 0, hence given the

knowledge of the distribution family, the variance will be used to determine the distribution uniquely.

Sometime it maybe extremely difficult to find an appropriate value for the error variance without examining the whole model carefully. For example, suppose $X_1, \cdots, X_n$ are direct causes of Y in the model, and $Y = f(X_1, \cdots, X_n) + \epsilon$, we may have no idea about how to specify the variance of ϵ so that it is not too large nor too small. In this case, we can specify a positive number r, which is defined as:

$$r = \frac{\text{Var}(\epsilon)}{\text{Var}(f(X_1, \cdots, X_n))}$$

- If a variable has direct causes in the model, we need to specify the functional relation between the variable and its causes. Here there are several factors we need to consider:
  - Suppose the variable Y is continuous, we need to specify whether the functional relation between the variable Y and its direct causes $X_1, \cdots, X_k$ can be represented as:

    $$Y = \sum_{i=1}^{k} f_i(X_i) + \epsilon$$

    or only as:

    $$Y = \sum_{i=1}^{k-1} f_i(X_i, X_{i+1}) + f_k(X_k, X_1) + \epsilon$$

    or only as:

    $$Y = f(X_1, \cdots, X_k) + \epsilon$$

    In the first case, we say that there is no interaction among the direct causes of Y. In the second case, we say that there are pairwise interactions among the direct causes of Y. In the third case, we say that there is a global interaction among the direct causes of Y.
  - If the variable Y is discrete, we need to specify the function that converts the function of its direct causes into a multinomial distribution. Here we have three choices:

    $$p_i = \frac{\exp\{f_i(x_1, \cdots, x_k)\}}{\sum_{j=1}^{m} \exp\{f_j(x_1, \cdots, x_k)\}}$$

    $$p_i = \frac{|f_i(x_1, \cdots, x_k)|}{\sum_{j=1}^{m} |f_j(x_1, \cdots, x_k)|}$$

$$p_i = \frac{(f_i(x_1, \cdots, x_k))^2}{\sum_{j=1}^{m}(f_j(x_1, \cdots, x_k))^2}$$

where $X_1, \cdots, X_k$ are direct causes of Y, m is the number of levels of Y, p_i is the probability of Y in the i^{th} level, and $f_i(x_1, \cdots, x_k)$ are the function of $X_1, \cdots, X_n$ used to compute the probability of Y in the i^{th} level.

– We also need to specify all of the function f_i's appearing above.

The follow categories of functions are allowed to be used in the GAM model:

Polynomial functions with degree up to 5, rational functions, logarithm function, exponential functions, and trigonometric function.

We will need to specify both the categories and the coefficients of the functions.

Sample Generator: The function of the sample generator should be clear from the previous discussion. Here we want to point out that the sample generator is not only used to generate data from a model, but also used to generate a random or partially random model. This is essential if we want to guarantee the quality of our randomly generated GAM Models.

To illustrate this point, consider a continuous variables Y and its direct causes $X_1, \cdots, X_k$. The functional relation between Y and $X_1, \cdots, X_k$ should be:

$$Y = f(X_1, \cdots, X_k) + \epsilon$$

where f may be a linear combination of functions allowed in a GAM model.

Now if the f is random chosen, it may happen that the influence of X_i on Y is ignorable. We definitely want to avoid this kind of situation. One way to solve the problem is to draw random functions iteratively:

1. First, randomly choose function f.
2. Generate data using the chosen function, check the ratio of the variance of $f(X_1, \cdots, X_k)$ to the variance of $f(X_1, \cdots, \mathrm{E}(X_i), \cdots, X_k)$ for each i.
3. If some ratio is extremely small, choose f again. Otherwise, stop.

Similarly, if Y is a discrete variable, we would like to use a similar strategy to choose functions of its direct causes such that the probability of Y in each of its levels is not extremely small.

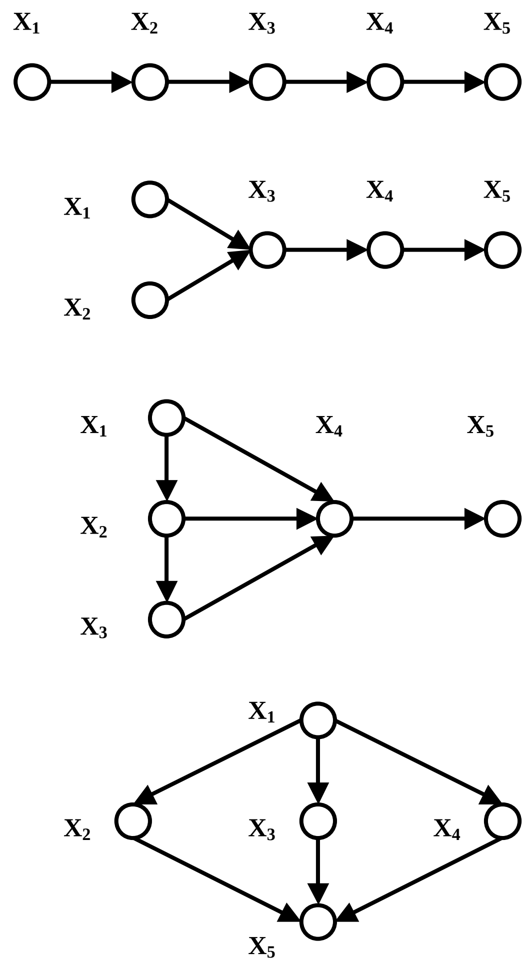

Figure 3: DAGs for Simulation Study

4.3 Graph

To test the GAM approach to non-parametric conditional independence test with AIC and BIC scores, we generated simulated samples based on 4 graphical structures.

All the samples are based on the 4 graphs shown in figure 3:

4.4 Samples

1. For the first graph, 150 continuous samples were generated: 50 of them with normal error, 50 with gamma, and 50 with mixture of normal error.

   Another 100 discrete samples were generated: 50 of them with 4-level discrete variables, and 50 with 8-level discrete variable.

   Another 50 mixture of continuous and discrete samples were generated: x_1, x_4: continuous; x_2, x_5:

discrete; x_3: could be either.

2. For the second graph, 300 continuous samples were generated: 150 of them with no interaction between parents, and 150 with interaction between parents. Among each group of 150 samples: 50 of them with normal error, 50 with gamma, and 50 with mixture of normal error.

   Another 100 discrete samples were generated: 50 of them with 4-level discrete variables, 50 with 8-level discrete variable.

   Another 50 mixture of continuous and discrete samples were generated: x_1, x_3, x_4: continuous, x_2, x_5: discrete.

3. For the third graph, 450 continuous samples were generated: 150 of them with no interaction among parents, 150 with pairwise interaction between parents, and 150 with global interaction among parents. Among each group of 150 samples: 50 of them with normal error, 50 with gamma, and 50 with mixture of normal error.

   Another 100 discrete samples were generated: 50 of them with 4-level discrete variables, 50 with 8-level discrete variable.

   Another 150 mixture of continuous and discrete samples were generated: x_1: continuous, x_3: discrete, x_4: continuous, x_2, x_5: could be either.

4. For the fourth graph, 450 continuous samples were generated: 150 of them with no interaction among parents, 150 with pairwise interaction between parents, and 150 with global interaction among parents. Among each group of 150 samples: 50 of them with normal error, 50 with gamma, and 50 with mixture of normal error.

5 Discussion

5.1 General Observations

By looking at the tables presented above, we can get the following observations:

1. It is easy to see that the GAM method should not be applied to discrete samples, for it will always choose the simpler regression model. For example, for the 100 discrete samples generated from the first DAG in figure 3, when testing whether $X_1 \perp\!\!\!\perp X_5 | X_3$, the GAM test using the BIC score said that X_1 and X_5 are independent given X_3 for all the 100 samples. This happens to be a coditional independence relation entailed by the DAG. However, when testing whether $X_1 \perp\!\!\!\perp X_3 | X_5$, the GAM test using the BIC score again claimed that X_1 and X_3 are independent given X_5 for all the 100 samples. Unfortunately, as a matter of fact, the DAG implies that X_1 and X_3 are *not* independent conditional on X_5.

   This tendency of choosing simpler regression model, hence the tendency of accepting the conditional independence hypothesis, is not surprising, for the GAM regression is not designed for all discrete predictors. When all predictors are discrete, the GAM regression will try to partition the sample into a contingency table according to the predictors, and try to regress the response variable on each cell of the contingency table.

2. Comparing the test result obtained with the AIC score and those with the BIC score, it is clear that the BIC score prefers simpler models, compared to the AIC score. For example, for the 149 continuous samples (each with sample size 500) generated from the first DAG in figure 3, when testing whether $X_1 \perp\!\!\!\perp X_5 | X_3$, the GAM test using the BIC score said that X_1 and X_5 are independent given X_3 for 132 samples, while the GAM test using the AIC score said that X_1 and X_5 are independent given X_3 for 83 samples. When testing whether $X_1 \perp\!\!\!\perp X_3 | X_5$, the GAM test using the BIC score again claimed that X_1 and X_3 are not independent given X_5 for 107 samples, while the GAM test using the AIC score said that X_1 and X_3 are not independent given X_5 for 122 samples.

   This is exactly what we expected, because the difference between AIC and BIC, up to a constant factor, is that the former has a penalty term equal to 2 times the equivalent degree of freedom, while the later has a penalty term equal to $\log(n)$ times the equivalent degree of freedom, with n being the samples size. Clearly, for any $n > e^2$, BIC will penalize model complexity more heavily than AIC.

3. In terms of the interaction among parents of a child variable, there is no significant difference in test accuracy among the three cases: no interaction, pairwise interaction, and global interaction.

For example, from the second DAG in figure 3, 50 continuous samples with no interaction among parents of child variables, and 50 continuous samples with global interactions among parents of child variables, were generated. Seven conditional independence tests were performed on these samples. For the samples allowing no global interaction, the average accuracy of conditional independence tests using the BIC score is approximately 0.69, while for the samples allowing global interactions, the average accuracy is approximately 0.70. This is more or less a surprise, considering the violation of the assumption of GAM model when pairwise or global interactions present in the samples.

One plausible explanation is that, although the additive assumption is violated, adding one more predictor that is not conditionally independent of the response variable given the current set of predictors can still further reduce the deviance of the fitted model. Another explanation is that our test performs poorly in cases where whether the parents are interacting with each other should make a difference, and the poor performance to some extent obscures the otherwise identifiable difference.

4. In terms of error distribution, we find that the accuracy of the GAM tests on the samples with mixture of normal error distribution usually is as good as, or even slightly better, than that of the GAM tests on the samples with norm error distriubtion. For example, for the 49 continuous samples (each with sample size 500) with normal error generated from the first DAG in figure 3, the average accuracy of 9 conditional independence tests with BIC score is 0.77, while the average accuracy of conditional independence tests on the 50 continuous samples with mixture of normal error generated from the same DAG is 0.81.

   One explanation of this result is that, because the mixture of normal error distribution in our simulation study is actually a mixture of two equally weighted normal distributions with identical variance but different mean, it is a symmetric error distribution. Hence, the GAM regression method works well.

   On the other hand, the behavior of the GAM test for the samples with *Gamma* error distribution does differ from that for the samples with symmetric error distribution, i.e., normal or mixture of normal. If we only look at the conditional independence test results for the samples with *Gamma* error, we could find that the GAM test is less prone to simpler models for samples with *Gamma* error. For example, for the 50 continuous samples with *Gamma* error generated from the first DAG in figure 3, using the BIC score, the average accuracy of the 5 tests where the true models are independent models is 0.74, while the average accuracy of the 4 tests where the true models are independent models is 0.79, while for the samples with normal error or mixture of normal error, using the BIC score, the average test accuracies are higher when the true models are independent models, and lower when the true models are dependent models. We believe this is because under the GAM model, the predictors cannot predict the response variable very well, so adding one more predictor usually can further reduce the deviance of the fitted model.

5. When using conditional independence test results to search for causal patterns, the performance is not satisfactory. To improve the the accuracy of conditional independence testing, we also tried the following two-way method, using the BIC score:

   Suppose we want to test whether X and Y are independent given $\boldsymbol{Z}$. We will apply GAM regression to the following 4 models:

$$M_1: \quad X = f(Y) + \sum_{i=1}^{k} g_i(Z_i) + \epsilon_1$$

$$M_2: \quad X = \sum_{i=1}^{k} h_i(Z_i) + \epsilon_2$$

$$M_3: \quad Y = f'(X) + \sum_{i=1}^{k} g_i'(Z_i) + \epsilon_1$$

$$M_4: \quad Y = \sum_{i=1}^{k} h_i'(Z_i) + \epsilon_2$$

   We will say that $X \perp\!\!\!\perp Y | \boldsymbol{Z}$ if and only if the BIC score of M_2 is less than or equal to that of M_1, *and* the BIC score of M_4 is less than or equal to that of M_3.

   Given that BIC score prefers the simpler model, this method should improve the performance of conditional independence test. However, from the simulation result, this method is not satisfying either. In general, the returned causal patterns contain more edges that are in the true causal pattern, but also contain more edges that are not in the true causal pattern. This is consistent with the fact that the two-way test method prefers the more complex model compared to the one-way method. The reason why this does not improve the test accuracy probably is because a correct causal pattern requires we get the correct conditional independence information in most cases.

5.2 AIC and BIC

To get some insight into the test results, it is helpful to take a closer look at the AIC and BIC score.

Recall that the formulae for AIC, BIC, and noise variance are:

$$AIC = RSS + 2\,df\,\sigma^2$$

$$BIC = RSS\,/\,\sigma^2 + \log(N)\,df$$

$$\sigma^2 = RSS_0\,/\,(N - df_0)$$

Substitute the formula for σ^2 to the formulae for AIC and BIC, we get:

$$AIC = RSS + 2\,df\frac{RSS_0}{N - df_0}$$

$$BIC = \frac{RSS}{RSS_0}(N - df_0) + \log(N)\,df$$

Suppose we are going to test whether a random variable X is independent of a random variable Y given a set of random variable $Z_1, \cdots, Z_k$. That is, we are going to test the two alternative models M_Z and $M_{Y,Z}$ as defined in equations 16) and 15. Let RSS_Z and df_Z be the residual sum square and degree of freedom for the model M_Z, and $RSS_{Y,Z}$ and $df_{Y,Z}$ for the model $M_{Y,Z}$. Then, using BIC, the test will support the hypothesis that X is independent of Y given $Z_1, \cdots, Z_k$ if and only if:

$$\frac{RSS_Z}{RSS_0}(N - df_0) + \log(N)\,df_Z < \frac{RSS_{Y,Z}}{RSS_0}(N - df_0) + \log(N)\,df_{Y,Z}$$
$$\iff \frac{RSS_Z - RSS_{Y,Z}}{RSS_0} < \frac{\log(N)}{N - df_0}(df_{Y,Z} - df_Z)$$

Similarly, if using AIC, the test will support the hypothesis that X is independent of Y given Z if and only if:

$$\frac{RSS_Z - RSS_{Y,Z}}{RSS_0} < \frac{2}{N - df_0}(df_{Y,Z} - df_Z)$$

From the above analysis, it is clear that, in this simulation study, when using the AIC and BIC scores to do conditional independence testing, what we really care about is the sign of the difference between $\frac{RSS_Z - RSS_{Y,Z}}{RSS_0}$, which is the ratio of the amount of reduced deviance by adding an extra predictor to the deviance of the most complicated model, and $\frac{2}{N - df_0}(df_{Y,Z} - df_Z)$ or $\frac{\log(N)}{N - df_0}(df_{Y,Z} - df_Z)$.In other words, we are not really making full use of the information provided by the AIC and BIC scores. This suggests that we might improve the test procedure by taking into account the actual value of the difference between the scores of two alternative models.

To illustrate the use of the actual value of the difference between the scores of two alternative models, suppose we want to use BIC score to test the alternative models M_Z and $M_{Y,Z}$. Let BIC_Z refer to the BIC score for model M_Z, $BIC_{Y,Z}$ refer to the BIC score for model $M_{Y,Z}$. Clearly, unless something wrong with the `gam` regression, we should have:

$$RSS_{Y,Z} < RSS_Z, \qquad df_{Y,Z} > df_Z$$

If we use the default smooth parameter when doing the GAM regression, in most tests, we should get: $df_{Y,Z} - df_Z \approx 4$

Combining the two above inequalities, we have:

$$BIC_{Y,Z} - BIC_Z < \log(N)[df_{Y,Z} - df_Z] \approx -4\log(N)$$

Note that there is virtually no lower bound for the above difference.

Now suppose that $X \perp\!\!\!\perp Y|(Z_1, \cdots, Z_k)$, then the difference between $BIC_{Y,Z}$ and BIC_Z should be positive, and we know that the difference cannot be greater than $\log(N)[df_{Y,Z} - df_Z]$. (Higher score means less preferred model.) Of course, if the test is inaccurate, we could end up with a negative difference. But usually this difference should be small.

On the other hand, if $X \not\perp\!\!\!\perp Y|(Z_1, \cdots, Z_k)$, then the different between $BIC_{Y,Z}$ and BIC_Z should be negative. Although quite often we find that this difference is positive, in some cases, we may get a negative difference with a large absolute value compared with $\log(N)[df_{Y,Z} - df_Z]$. In this case, we are pretty sure that the dependence relation is true.

5.3 Beyond AIC and BIC

The GAM regression actually provides far more information than the AIC and BIC scores. Some information may help to identify situations where the GAM model is not applicable.

One way to see whether a GAM regression model is applicable to some sample is to look at the degrees of freedom of the GAM regression model. Usually a large number of degrees of freedom of a regression model suggests that this model is not a good model of the sample data. In our simulation study, we find that, for a sample with *Gamma* error, the GAM regression tends to return a regression model with unexpected large degrees of freedom. For example, consider the 149 continuous samples (each with sample size 500) generated from the first DAG in figure 3. When we regress X_2 on X_3 and X_4, because the default smoothing parameter for the spline smoother used in the GAM regression is roughly corresponding to 4 degrees of freedom, the regression model for each sample should have roughly 9 degrees of freedom: 4 for each predictor, and 1 for a constant additive term. This is the case for the 99 samples with normal error or mixture of normal error. The degrees of freedom of the regression models for these samples are all between 8.5 and 9.5. However, for the 50 samples with *Gamma* error, the degrees of freedom of the regression model for 31 of them are greater than 9.5—some models have more than 100 degrees of freedom.

Some information provided by the GAM regression can be used to predict the reliability of the GAM conditional independence test based on AIC or BIC scores.

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 residual sum square of the model $X \sim 1$, *noise deviance* be the residual sum square of the model $X \sim f(Y) + \sum_{i=1}^{k} g_i(Z_i) + \sum_{j=1}^{m} h_j(U_j)$, *independent deviance* be the residual sum square of the model $X \sim \sum_{i=1}^{k} l_i(Z_i)$, and *dependent deviance* be the residual sum square of the model $X \sim s(Y) + \sum_{i=1}^{k} t_i(Z_i)$. It turns out that the 3 ratios of *noise deviance*, *independent deviance*, and *dependent deviance*, to the *null deviance*, can help us to estimate the sensitivity of the GAM conditional independence test to small perturbation in the GAM regression model, hence can be used to estimate the reliability of the conditional independence test.For example, when using the BIC score, we could use the following heuristic rules:

1. If the *noise deviance* is very small compared to the *null deviance*, and the *independent deviance* is not very close to the *null deviance*, then if the result is dependence, it is a **weak dependence**, and if the result is independence, it is a **strong independence**.

   Justification: a small *noise deviance* implies that the GAM test is too sensitive to small difference in residual sum squares, hence tends to choose the more complex model.

2. If the *noise deviance* is relatively large compared to the *null deviance*, and the *independent deviance* is not very close to the *null deviance*, then if the result is dependence, it is a **strong dependence**, and if the result is independence, it is a **weak independence**.

   Justification: a large *noise deviance* implies that the GAM test is too insensitive to small difference in residual sum squares, hence tends to choose the simpler model.

3. If the *independent deviance* and the *dependent deviance* are very close to the *noise deviance*, then if the result is dependence, it is a **weak dependence**, and if the result is independence, it is a **weak independence**.

Justification: If two alternative models both have residual sum squares almost as small as the most complicated model, the GAM test really could not tell which model is better.

To see how these rules work, we could apply them to one of the most problematic test results in our simulation study: the test of whether $X_2 \perp\!\!\!\perp X_4|X_3$ using the BIC score, where the true model is $X_2 \perp\!\!\!\perp X_4|X_3$. Consider the 49 continuous samples (each with sample size 500) with normal error generated from the first DAG in figure 3. The GAM tests using the BIC score have an accuracy of 0.76, compared to an average accuracy of 0.90 for the GAM tests using the BIC score for these samples on the other 4 conditional independence hypotheses, where the true hypotheses are the conditional independence hypotheses. If we take a closer look at the GAM regression results for the 12 samples where the GAM test fails, we find for 10 of them, the ratio of the *noise deviances* to the *null deviance* is less than 0.12; and the for another 2 of them the ratio of the *independent deviance* to the *null deviance* is greater than 0.95. Therefore, according to the first and the third heuristic rules, for all of these 12 samples, the resulting dependence relations are weak dependence.

Moreover, if we look at the GAM regression results for the 37 samples where the GAM tests are correct, according to our heuristics, we are confident about 28 of them. This implies that, if we restrict ourselves only to those test results about which we are confident, the GAM test procedure is useful for more than half of the times.

6 Future Work

If we want to further improve the accuracy of conditional independence test, we might want to consider the following options:

1. Make the distribution of the `gam` function more adaptive. Currently, for continuous data, we use `gaussian`. Unfortunately, while other distributions are allowed by `gam`, none of them looks like a better choice.

2. Change the choice of smoothing parameter, or, more precisely, the choice of degree of freedom for spline smoothing. Currently we use the default value,which is 4. But we might try to find the optimal values for each predictor in the GAM regression model. One problem with this approach is that choosing the smoothing parameter adaptively is usually compuationally expensive.

3. Change the way to measure the model complexity. Currently, we use one definition of the equivalent degrees of freedom. We could try other ways of measuring the model complexity. For example, we could use the generalized degrees of freedom.[14]

A more promising approach, still based on GAM method, is to train some decision trees with much more information than we currently used to get the sign of the difference in AIC or BIC scores. For example, to test whether X and Y are independent give Z, we may first train a decision tree with the differences of the BIC scores for the 2 alternative regression models, *null deviance*, *noise deviance*, *independent deviance*, and *dependent deviance*. (See definitions of these deviances in section 5.3.) Then we train another decision tree to estimate how confident we are about the results given by the the first tree.

An alternative to the constraint based search is the use of maximum likelihood search. Given a DAG, assuming causal sufficiency and normal error term, the GAM approach can provide, for each random variable X, an estimate of its log-likelihood:

$$-\frac{RSS}{2\sigma^2} - \log(\sigma) + C$$

Thus, for a set of m random variables $\boldsymbol{X} = \{X_1, \cdots, X_m\}$, we could choose a DAG G that minimizes:

$$R_1' = \sum_{i=1}^{m} \left[\frac{RSS_{X_i|Par(X_i|G)}}{\sigma^2_{X_i|Par(X_i|G)}} + \log(\sigma^2_{X_i|Par(X_i|G)}) + \log(N)\, df_{X_i|Par(X_i|G)} \right]$$

where $Par_{X_i|G}$ is the set of variables that are parents of X_i in G.

Equivalently, we can choose a G that minimizes:

[14] Ye (1999).

$$R_1 = \sum_{i=1}^{m} \left[\log \left(\frac{RSS_{X_i|Par(X_i|G)}}{N - df_{X_i|Par(X_i|G)}} \right) + (\log(N) - 1)\, df_{X_i|Par(X_i|G)} \right]$$

Note that here we do not estimate the noise variance for each variable X_i based on a model that takes $\{X_i\}^c$ as predictors. Rather, we estimate the noise variance based on graph G.

An improvement to the above formula is to replace the log-likelihood of each exogenous variable with the value obtained from density estimation. For example, suppose we use kernel density estimation and some bandwidth selection criterion. For each exogenous variable X_i, let f_i be the estimated density function. Then we would like to choose a G that minimizes:

$$R_2 = \sum_{\{i:|Par(X_i|G)|>0\}} \left[\log \left(\frac{RSS_{X_i|Par(X_i|G)}}{N - df_{X_i|Par(X_i|G)}} \right) + (\log(N) - 1)\, df_{X_i|Par(X_i|G)} \right] + \sum_{\{i:|Par(X_i|G)|=0\}} \log(f_i(X_i))$$

The new formula essentially removes the assumption that all exogenous variables have normal distribution.

To minimize R_1 or R_2, we could either apply GAM regression with a fixed degree of freedom, say, 4. Or we could use some criterion to find optimal degree of freedoms for each regression. Obviously, the latter will be computationally more expensive.

Also, it seems that the GAM method is sensitive to the direction of the regression. This is true especially if the functional relation between parents and child is non-monotone on a rectangle with significant measure. Therefore, we may expect that the maximum likelihood approach should have a good chance to pick up a model close to the true model. We can also use R_2 to compare two statistically indistinguishable models.

Finally, one reason why we choose GAM over density estimation is the that the later method is subject to the curse of dimensionality, hence is not applicable for high dimensional test. However, our simulation result suggests that because of its poor accuracy, application of GAM method should also be restricted to low dimensional tests. Therefore, we may want to explore the low dimension test with density estimation, and see whether this will be a better approach.

7 References

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.

George, E. (2000) “The Variable Selection Problem”.

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.

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”, *Annals of Statistics Vol. 6*, pp.461-464.

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

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.

8 Appendix: Summary of Result

8.1 How to read the tables:

test The true model.

setting The types of the samples.

continuous/nominal/continuous and nominal The sample consists of only continuous/nominal discrete/mixture of continuous and nomimal discrete data.

normal/gamma/mix-normal error The error terms and exogenous variables in the sample have normal/gamma/mix-normal distributions.

4/8 level For discrete samples, each variable has 4/8 distinct levels.

aic Results of conditional independence test using the AIC score. The left column gives the number of samples where the simpler model were accepted. The right column gives the portion of samples where the true model were accepted. (Note that some times the true model was the simpler model, sometime the true model was the more complicated one.)

bic Results of conditional independence test using the BIC score. The left column gives the number of samples where the simpler model were accepted. The right column gives the portion of samples where the true model were accepted. (Note that some times the true model was the simpler model, sometime the true model was the more complicated one.)

# of samples Number of samples (of the type given by the **setting** column) for each conditional independence test.

8.2 First Graph—v5n1

Test for samples of size 2000

test	setting	aic		bic		# of samples
$x_1 \perp\!\!\!\perp x_5 \| x_3$	continuous, normal error	28	0.60	42	0.89	47
$x_1 \not\perp\!\!\!\perp x_3 \| x_5$	continuous, normal error	6	0.87	11	0.77	47
$x_1 \perp\!\!\!\perp x_5 \| x_4$	continuous, normal error	23	0.49	39	0.83	47
$x_1 \not\perp\!\!\!\perp x_4 \| x_5$	continuous, normal error	10	0.79	20	0.57	47
$x_1 \not\perp\!\!\!\perp x_5$	continuous, normal error	9	0.81	22	0.53	47
$x_1 \not\perp\!\!\!\perp x_2$	continuous, normal error	2	0.96	3	0.94	47

test	setting	aic		bic		# of samples
$x_1 \perp\!\!\!\perp x_5 \mid x_3$	continuous, normal error	27	0.55	46	0.94	49
	continuous, gamma error	27	0.54	39	0.78	50
	continuous, mix-normal error	29	0.58	47	0.94	50
	nominal, 4 levels	50	1.00	50	1.00	50
	nominal, 8 levels	50	1.00	50	1.00	50
	continuous and nominal	44	0.88	50	1.00	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_3 \mid x_5$	continuous, normal error	10	0.80	17	0.65	49
	continuous, gamma error	8	0.84	13	0.74	50
	continuous, mix-normal error	9	0.82	12	0.76	50
	nominal, 4 levels	49	0.02	50	0.00	50
	nominal, 8 levels	47	0.06	50	0.00	50
	continuous and nominal	15	0.70	21	0.58	50

test	setting	aic		bic		# of samples
$x_5 \perp\!\!\!\perp x_1 \mid x_3$	continuous, normal error	24	0.49	42	0.86	49
	continuous, gamma error	24	0.48	35	0.70	50
	continuous, mix-normal error	25	0.50	47	0.94	50
	nominal, 4 levels	47	0.94	50	1.00	50
	nominal, 8 levels	35	0.70	48	0.96	50
	continuous and nominal	39	0.78	49	0.98	50

test	setting	aic		bic		# of samples
$x_4 \perp\!\!\!\perp x_2 \mid x_3$	continuous, normal error	33	0.67	44	0.90	49
	continuous, gamma error	35	0.70	43	0.86	50
	continuous, mix-normal error	34	0.68	45	0.90	50
	nominal, 4 levels	46	0.92	48	0.96	50
	nominal, 8 levels	47	0.94	50	1.00	50
	continuous and nominal	45	0.90	49	0.98	50

test	setting	aic		bic		# of samples
$x_4 \not\perp\!\!\!\perp x_3 \mid x_2$	continuous, normal error	9	0.82	13	0.73	49
	continuous, gamma error	8	0.84	11	0.78	50
	continuous, mix-normal error	4	0.92	7	0.86	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_5$	continuous, normal error	7	0.86	27	0.45	49
	continuous, gamma error	3	0.94	9	0.82	50
	continuous, mix-normal error	9	0.82	19	0.62	50

test	setting	aic		bic		# of samples
$x_2 \perp\!\!\!\perp x_4 \mid x_3$	continuous, normal error	22	0.45	37	0.76	49
	continuous, gamma error	23	0.46	33	0.66	50
	continuous, mix-normal error	28	0.56	39	0.78	50
	nominal, 4 levels	49	0.98	50	1.00	50
	nominal, 8 levels	36	0.72	47	0.94	50
	continuous and nominal	37	0.74	48	0.96	50

test	setting	aic		bic		# of samples
$x_1 \perp\!\!\!\perp x_5 \mid x_4$	continuous, normal error	19	0.39	43	0.88	49
	continuous, gamma error	22	0.44	35	0.70	50
	continuous, mix-normal error	24	0.48	42	0.84	50
	continuous and nominal	43	0.86	50	1.00	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_4 \mid x_5$	continuous, normal error	14	0.71	25	0.49	49
	continuous, gamma error	10	0.80	14	0.72	50
	continuous, mix-normal error	13	0.74	17	0.66	50
	continuous and nominal	21	0.58	36	0.28	50

8.3 Second Graph—v5n3

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_2 \mid x_3$	continuous, normal error, no interaction	16	0.68	26	0.48	50
	continuous, gamma error, no interaction	9	0.82	17	0.66	50
	continuous, mix-normal error, no interaction	11	0.78	20	0.60	50
	continuous, normal error, global interaction	14	0.72	24	0.52	50
	continuous, gamma error, global interaction	9	0.82	16	0.68	50
	continuous, mix-normal error, global interaction	10	0.80	14	0.72	50
	nominal, 4 levels	49	0.02	50	0.00	50
	nominal, 8 levels	50	0.00	50	0.00	50
	continuous and nominal	5	0.90	10	0.80	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_2 \mid x_5$	continuous, normal error, no interaction	19	0.62	35	0.30	50
	continuous, gamma error, no interaction	15	0.70	30	0.40	50
	continuous, mix-normal error, no interaction	16	0.68	37	0.26	50
	continuous, normal error, global interaction	25	0.50	34	0.32	50
	continuous, gamma error, global interaction	14	0.72	24	0.52	50
	continuous, mix-normal error, global interaction	19	0.62	35	0.30	50
	nominal, 4 levels	50	0.00	50	0.00	50
	nominal, 8 levels	50	0.00	50	0.00	50
	continuous and nominal	19	0.62	33	0.34	50

test	setting	aic		bic		# of samples
$x_1 \perp\!\!\!\perp x_5 \mid x_3$	continuous, normal error, no interaction	37	0.74	49	0.98	50
	continuous, gamma error, no interaction	38	0.76	49	0.98	50
	continuous, mix-normal error, no interaction	37	0.74	50	1.00	50
	continuous, normal error, global interaction	45	0.90	50	1.00	50
	continuous, gamma error, global interaction	29	0.58	45	0.90	50
	continuous, mix-normal error, global interaction	36	0.72	49	0.98	50
	continuous and nominal	34	0.68	49	0.98	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_3 \mid x_5$	continuous, normal error, no interaction	8	0.84	15	0.70	50
	continuous, gamma error, no interaction	8	0.84	17	0.66	50
	continuous, mix-normal error, no interaction	8	0.84	14	0.72	50
	continuous, normal error, global interaction	9	0.82	21	0.58	50
	continuous, gamma error, global interaction	7	0.86	16	0.68	50
	continuous, mix-normal error, global interaction	7	0.86	16	0.68	50

test	setting	aic		bic		# of samples
$x_5 \perp\!\!\!\perp x_1 \mid x_3$	continuous, normal error, no interaction	26	0.52	33	0.66	50
	continuous, gamma error, no interaction	18	0.36	27	0.54	50
	continuous, mix-normal error, no interaction	27	0.54	42	0.84	50
	continuous, normal error, global interaction	26	0.52	41	0.82	50
	continuous, gamma error, global interaction	22	0.44	33	0.66	50
	continuous, mix-normal error, global interaction	23	0.46	40	0.80	50
	continuous and nominal	43	0.86	49	0.98	50

test	setting	aic		bic		# of samples
$x_5 \not\perp\!\!\!\perp x_3 \vert x_1$	continuous, normal error, no interaction	12	0.76	16	0.68	50
	continuous, gamma error, no interaction	3	0.94	7	0.86	50
	continuous, mix-normal error, no interaction	13	0.74	16	0.68	50
	continuous, normal error, global interaction	11	0.78	14	0.72	50
	continuous, gamma error, global interaction	5	0.90	7	0.86	50
	continuous, mix-normal error, global interaction	15	0.70	18	0.64	50

test	setting	aic		bic		# of samples
$x_1 \perp\!\!\!\perp x_2$	continuous, normal error, no interaction	21	0.42	37	0.74	50
	continuous, gamma error, no interaction	15	0.30	42	0.84	50
	continuous, mix-normal error, no interaction	17	0.34	42	0.84	50
	continuous, normal error, global interaction	27	0.54	42	0.84	50
	continuous, gamma error, global interaction	16	0.32	30	0.60	50
	continuous, mix-normal error, global interaction	21	0.42	39	0.78	50

test	setting	aic		bic		# of samples
$x_2 \not\perp\!\!\!\perp x_1 \vert x_3$	continuous and nominal	22	0.56	41	0.18	50
$x_2 \not\perp\!\!\!\perp x_1 \vert x_5$	continuous and nominal	42	0.16	49	0.02	50

8.4 Third Graph—v5n2

test	setting	aic		bic		# of samples
$x_1 \perp\!\!\!\perp x_3 \vert x_2$	continuous, normal error, no interaction	34	0.68	46	0.92	50
	continuous, gamma error, no interaction	31	0.62	43	0.86	50
	continuous, mix-normal error, no interaction	41	0.82	48	0.96	50
	continuous, normal error, pairwise interaction	37	0.74	49	0.98	50
	continuous, gamma error, pairwise interaction	38	0.76	45	0.90	50
	continuous, mix-normal error, pairwise interaction	40	0.80	49	0.98	50
	continuous, normal error, global interaction	38	0.76	49	0.98	50
	continuous, gamma error, global interaction	29	0.58	41	0.82	50
	continuous, mix-normal error, global interaction	43	0.86	50	1.00	50
	nominal, 4 levels	50	1.00	50	1.00	50
	nominal, 8 levels	50	1.00	50	1.00	50
	continuous and nominal, no interaction	34	0.68	46	0.92	50
	continuous and nominal, pairwise interaction	40	0.80	48	0.96	50
	continuous and nominal, global interaction	32	0.64	46	0.92	50

test	setting	aic		bic		# of samples
$x_3 \perp\!\!\!\perp x_1 \vert x_2$	continuous, normal error, no interaction	35	0.70	46	0.92	50
	continuous, gamma error, no interaction	28	0.56	35	0.70	50
	continuous, mix-normal error, no interaction	40	0.80	47	0.94	50
	continuous, normal error, pairwise interaction	34	0.68	43	0.86	50
	continuous, gamma error, pairwise interaction	38	0.76	45	0.90	50
	continuous, mix-normal error, pairwise interaction	43	0.86	48	0.96	50
	continuous, normal error, global interaction	42	0.84	47	0.94	50
	continuous, gamma error, global interaction	32	0.64	40	0.80	50
	continuous, mix-normal error, global interaction	43	0.86	49	0.98	50
	nominal, 4 levels	48	0.96	50	1.00	50
	nominal, 8 levels	48	0.96	48	0.96	50
	continuous and nominal, no interaction	40	0.80	47	0.94	50
	continuous and nominal, pairwise interaction	36	0.72	47	0.94	50
	continuous and nominal, global interaction	28	0.56	46	0.92	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_3 \| x_2, x_4$	continuous, normal error, no interaction	29	0.42	40	0.20	50
	continuous, gamma error, no interaction	24	0.52	37	0.26	50
	continuous, mix-normal error, no interaction	36	0.28	41	0.18	50
	continuous, normal error, pairwise interaction	31	0.38	44	0.12	50
	continuous, gamma error, pairwise interaction	35	0.30	42	0.16	50
	continuous, mix-normal error, pairwise interaction	35	0.30	47	0.06	50
	continuous, normal error, global interaction	32	0.36	43	0.14	50
	continuous, gamma error, global interaction	21	0.58	34	0.32	50
	continuous, mix-normal error, global interaction	38	0.24	44	0.12	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_3 \| x_2, x_5$	continuous, normal error, no interaction	33	0.34	44	0.12	50
	continuous, gamma error, no interaction	33	0.34	41	0.18	50
	continuous, mix-normal error, no interaction	37	0.26	46	0.08	50
	continuous, normal error, pairwise interaction	33	0.34	47	0.06	50
	continuous, gamma error, pairwise interaction	33	0.34	43	0.14	50
	continuous, mix-normal error, pairwise interaction	38	0.24	49	0.02	50
	continuous, normal error, global interaction	33	0.34	47	0.06	50
	continuous, gamma error, global interaction	26	0.48	36	0.28	50
	continuous, mix-normal error, global interaction	41	0.18	47	0.06	50
	continuous and nominal, no interaction	23	0.54	36	0.28	50
	continuous and nominal, pairwise interaction	34	0.32	42	0.16	50
	continuous and nominal, global interaction	28	0.44	39	0.22	50

test	setting	aic		bic		# of samples
$x_3 \not\perp\!\!\!\perp x_1 \| x_2, x_5$	continuous, normal error, no interaction	32	0.36	45	0.10	50
	continuous, gamma error, no interaction	25	0.50	34	0.32	50
	continuous, mix-normal error, no interaction	35	0.30	45	0.10	50
	continuous, normal error, pairwise interaction	35	0.30	44	0.12	50
	continuous, gamma error, pairwise interaction	33	0.34	43	0.14	50
	continuous, mix-normal error, pairwise interaction	41	0.18	48	0.04	50
	continuous, normal error, global interaction	40	0.20	47	0.06	50
	continuous, gamma error, global interaction	26	0.48	39	0.22	50
	continuous, mix-normal error, global interaction	40	0.20	46	0.08	50
	continuous and nominal, no interaction	38	0.24	46	0.08	50
	continuous and nominal, pairwise interaction	32	0.36	47	0.06	50
	continuous and nominal, global interaction	25	0.50	43	0.14	50

test	setting	aic		bic		# of samples
$x_3 \perp\!\!\!\perp x_5 \| x_4$	continuous, normal error, no interaction	13	0.26	27	0.54	50
	continuous, gamma error, no interaction	16	0.32	28	0.56	50
	continuous, mix-normal error, no interaction	11	0.22	29	0.58	50
	continuous, normal error, pairwise interaction	20	0.40	32	0.64	50
	continuous, gamma error, pairwise interaction	13	0.26	21	0.42	50
	continuous, mix-normal error, pairwise interaction	17	0.34	35	0.70	50
	continuous, normal error, global interaction	19	0.38	36	0.72	50
	continuous, gamma error, global interaction	18	0.36	30	0.30	50
	continuous, mix-normal error, global interaction	16	0.32	27	0.54	50
	continuous and nominal, no interaction	38	0.76	48	0.96	50
	continuous and nominal, pairwise interaction	41	0.82	49	0.98	50
	continuous and nominal, global interaction	34	0.68	45	0.90	50

test	setting	aic		bic		# of samples
$x_5 \perp\!\!\!\perp x_3 \| x_4$	continuous, normal error, no interaction	46	0.92	50	1.00	50
	continuous, gamma error, no interaction	39	0.78	47	0.94	50
	continuous, mix-normal error, no interaction	38	0.76	48	0.96	50
	continuous, normal error, pairwise interaction	44	0.88	50	1.00	50
	continuous, gamma error, pairwise interaction	36	0.72	45	0.90	50
	continuous, mix-normal error, pairwise interaction	42	0.84	50	1.00	50
	continuous, normal error, global interaction	43	0.86	50	1.00	50
	continuous, gamma error, global interaction	42	0.84	48	0.96	50
	continuous, mix-normal error, global interaction	45	0.90	50	1.00	50
	continuous and nominal, no interaction	46	0.92	50	1.00	50
	continuous and nominal, pairwise interaction	45	0.90	49	0.98	50
	continuous and nominal, global interaction	45	0.90	48	0.96	50

test	setting	aic		bic		# of samples
$x_5 \not\perp\!\!\!\perp x_4$	nominal, 4 levels	46	0.08	50	0.00	50
	nominal, 8 levels	45	0.10	48	0.04	50

8.5 Fourth Graph—v5n6

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_5 \| x_2$	continuous, normal error, no interaction	5	0.90	8	0.82	50
	continuous, gamma error, no interaction	2	0.96	10	0.80	50
	continuous, mix-normal error, no interaction	6	0.88	18	0.64	50
	continuous, normal error, pairwise interaction	9	0.82	23	0.54	50
	continuous, gamma error, pairwise interaction	5	0.90	16	0.68	50
	continuous, mix-normal error, pairwise interaction	2	0.96	13	0.74	50
	continuous, normal error, global interaction	9	0.82	21	0.58	50
	continuous, gamma error, global interaction	5	0.90	12	0.76	50
	continuous, mix-normal error, global interaction	6	0.88	23	0.54	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_5 \| x_2, x_3$	continuous, normal error, no interaction	13	0.74	27	0.46	50
	continuous, gamma error, no interaction	14	0.72	25	0.50	50
	continuous, mix-normal error, no interaction	16	0.68	34	0.32	50
	continuous, normal error, pairwise interaction	20	0.60	34	0.32	50
	continuous, gamma error, pairwise interaction	17	0.66	31	0.38	50
	continuous, mix-normal error, pairwise interaction	15	0.70	33	0.34	50
	continuous, normal error, global interaction	20	0.60	38	0.24	50
	continuous, gamma error, global interaction	9	0.82	20	0.60	50
	continuous, mix-normal error, global interaction	15	0.70	32	0.36	50

test	setting	aic		bic		# of samples
$x_1 \perp\!\!\!\perp x_5$ $\| x_2, x_3, x_4$	continuous, normal error, no interaction	36	0.72	46	0.92	50
	continuous, gamma error, no interaction	33	0.66	44	0.88	50
	continuous, mix-normal error, no interaction	33	0.66	45	0.90	50
	continuous, normal error, pairwise interaction	28	0.56	47	0.94	50
	continuous, gamma error, pairwise interaction	27	0.54	42	0.84	50
	continuous, mix-normal error, pairwise interaction	37	0.74	48	0.96	50
	continuous, normal error, global interaction	34	0.68	48	0.96	50
	continuous, gamma error, global interaction	27	0.54	39	0.78	50
	continuous, mix-normal error, global interaction	34	0.68	50	1.00	50

test	setting	aic		bic		# of samples
$x_2 \perp\!\!\!\perp x_3 \mid x_1$	continuous, normal error, no interaction	33	0.66	42	0.84	50
	continuous, gamma error, no interaction	34	0.68	39	0.78	50
	continuous, mix-normal error, no interaction	35	0.70	41	0.82	50
	continuous, normal error, pairwise interaction	29	0.58	37	0.74	50
	continuous, gamma error, pairwise interaction	28	0.56	37	0.74	50
	continuous, mix-normal error, pairwise interaction	34	0.68	41	0.82	50
	continuous, normal error, global interaction	37	0.74	39	0.78	50
	continuous, gamma error, global interaction	29	0.58	39	0.78	50
	continuous, mix-normal error, global interaction	34	0.68	38	0.76	50

test	setting	aic		bic		# of samples
$x_2 \not\perp\!\!\!\perp x_3 \mid x_1, x_5$	continuous, normal error, no interaction	25	0.50	37	0.26	50
	continuous, gamma error, no interaction	33	0.34	38	0.24	50
	continuous, mix-normal error, no interaction	25	0.50	35	0.30	50
	continuous, normal error, pairwise interaction	28	0.44	35	0.30	50
	continuous, gamma error, pairwise interaction	27	0.46	37	0.26	50
	continuous, mix-normal error, pairwise interaction	31	0.38	38	0.24	50
	continuous, normal error, global interaction	31	0.38	39	0.22	50
	continuous, gamma error, global interaction	24	0.52	37	0.26	50
	continuous, mix-normal error, global interaction	27	0.46	35	0.30	50

test	setting	aic		bic		# of samples
$x_1 \not\perp\!\!\!\perp x_2 \mid x_3, x_4$	continuous, normal error, no interaction	2	0.96	9	0.82	50
	continuous, gamma error, no interaction	9	0.82	13	0.74	50
	continuous, mix-normal error, no interaction	5	0.90	10	0.80	50
	continuous, normal error, pairwise interaction	4	0.92	13	0.74	50
	continuous, gamma error, pairwise interaction	6	0.88	12	0.76	50
	continuous, mix-normal error, pairwise interaction	2	0.96	10	0.80	50
	continuous, normal error, global interaction	6	0.88	12	0.76	50
	continuous, gamma error, global interaction	9	0.82	18	0.64	50
	continuous, mix-normal error, global interaction	4	0.92	13	0.74	50