Journal of Machine Learning Research 9 (2008) 967-991 Submitted 9/06; Revised 9/07; Published 5/08

Search for Additive Nonlinear Time Series Causal Models

Tianjiao Chu TIC19@PITT.EDU
Department of Obstetrics, Gynecology & Reproductive Sciences
University of Pittsburgh
204 Craft Ave., Room B409
Pittsburgh, PA 15213, USA

Clark Glymour CG09@ANDREW.CMU.EDU
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Greg Ridgeway

Abstract

Pointwise consistent, feasible procedures for estimating contemporaneous linear causal structure from time series data have been developed using multiple conditional independence tests, but no such procedures are available for non-linear systems. We describe a feasible procedure for learning a class of non-linear time series structures, which we call additive non-linear time series. We show that for data generated from stationary models of this type, two classes of conditional independence relations among time series variables and their lags can be tested efficiently and consistently using tests based on additive model regression. Combining results of statistical tests for these two classes of conditional independence relations and the temporal structure of time series data, a new consistent model specification procedure is able to extract relatively detailed causal information. We investigate the finite sample behavior of the procedure through simulation, and illustrate the application of this method through analysis of the possible causal connections among four ocean indices. Several variants of the procedure are also discussed.

Keywords: conditional independence test, contemporaneous causation, additive model regression, Granger causality, ocean indices

1. Introduction

For stationary time series of four or more dimensions, Swanson and Granger (1997) proposed to determine contemporaneous causation—causal influences occurring more rapidly than the sampling interval of the time series data—by regressing each time series variable on all lags of all variables considered and using the residuals to test for vanishing partial correlations. Using search procedures for directed acyclic graphical linear models, in particular, the PC algorithm (Spirtes et al., 2000), Bessler et al. (2002), Demiralp and Hoover (2003), and Hoover (2005) generalized Swanson and Granger's procedure to allow specification searches for contemporaneous linear systems among all partial orderings of the dependencies among the variables. Moneta (2003) derived the correction needed for the fact that the correlations are obtained from residuals of a regression, and applied it to a set of cointegrated variables.

All these methods are designed for linear systems with joint Normal distributions, and allow neither unrecorded (latent) common causes nor feedbacks. One source of these limitations is the

search algorithm used by all of these procedures, PC, which is known to be consistent only in the absence of feedback relations and latent common causes. In principle, some of these difficulties can be met by replacing PC with related algorithms: the FCI algorithm (Spirtes et al., 2000), which allows latent variables, or an algorithm due to Richardson and Spirtes (1999) that allows linear feedback relations, though no algorithm is available that is consistent for search for linear causal models when both latent variables *and* feedback may be present.

More fundamentally, PC and related algorithms require conditional independence information about the random variables as input, and are therefore limited to distribution families for which conditional independence tests of arbitrary order are available, such as Multinomial and Normal distributions. (Another group of causal inference algorithms that are based on model scores, such as Bayesian posteriors, are unable to handle either latent variables *or* feedbacks, except under extra constraints (Silva et al., 2006; Drton et al., 2006). For *non-Gaussian* linear models with latent variables, independent component analysis based algorithms (Hoyer et al., 2006) could be more informative than PC and FCI.) Extending the PC and related algorithms based on conditional independence constraints to a larger class of systems that includes nonlinear continuous models requires more general conditional independence tests. We begin by considering some of the difficulties involved with finding such tests.

In theory, using nonparametric density estimation, we can test conditional independence for any set of random variables which have a joint density with respect to the Lebesgue measure. For example, let the joint density of $\{X,Y,Z\}$ be $f_{XYZ}(x,y,z)$, the joint density of $\{X,Z\}$ be $f_{XZ}(x,z)$, the joint density of $\{Y,Z\}$ be $f_{YZ}(y,z)$, and the marginal density of Z be $f_Z(z)$. We could test if X and Y are independent given Z by testing if the Hellinger distance between $f_{XYZ}(x,y,z)f_Z(z)$ and $f_{XZ}(x,z)f_{YZ}(y,z)$ is 0. For example, Su and White (2007) propose a conditional independence test for stationary time series satisfying certain conditions, based on a weighted Hellinger distance between $f_{X|YZ}(x;y,z)$ and $f_{X|Z}(x;z)$, where $f_{X|YZ}(x;y,z)$ and $f_{X|Z}(x;z)$ are densities of the conditional distributions of X given $\{Y,Z\}$ and Z respectively. However, this approach requires nonparametric density estimation of multivariate distributions, which is subject to the curse of dimensionality: as the number of variables increases, the data points become sparse rapidly in the space spanned by the variables.

Baek and Brock (1992) and Hiemstra and Jones (1994) proposed a nonparametric method intended for Granger causality testing of nonlinear time series. Consider a bivariate time series $\{X_t,Y_t\}$, $t = 1,\cdots$, let $\boldsymbol{X}_t^m = (X_t,\cdots,X_{t+m-1})$ for some m, they proposed to test if $\boldsymbol{X}_t^m$ and $\boldsymbol{Y}_{t-b}^b$ are independent given $\boldsymbol{X}_{t-a}^a$ by testing the following null hypothesis:

$$P\left(\|\boldsymbol{X}_t^m - \boldsymbol{X}_s^m\|_\infty < e \mid \|\boldsymbol{X}_{t-a}^a - \boldsymbol{X}_{s-a}^a\|_\infty < e, \|\boldsymbol{Y}_{t-b}^b - \boldsymbol{Y}_{s-b}^b\|_\infty < e\right)$$
$$= P\left(\|\boldsymbol{X}_t^m - \boldsymbol{X}_s^m\|_\infty < e \mid \|\boldsymbol{X}_{t-a}^a - \boldsymbol{X}_{s-a}^a\|_\infty < e\right).$$

Unfortunately, only under some specific conditions is the above null hypothesis equivalent to the hypothesis that $\boldsymbol{X}_t^m$ is independent of $\boldsymbol{Y}_{t-b}^b$ given $\boldsymbol{X}_{t-a}^a$ (Diks and Panchenko, 2006).

Bell et al. (1996) considered additive model regression (Hastie and Tibshirani, 1990) for conditional independence tests in their study of nonlinear Granger causality. An additive model assumes that the response variable Y is a linear combination of univariate smooth functions of predictors $\boldsymbol{X} = \{X_1,\cdots,X_p\}$ plus an independent error term. That is:

$$Y = \sum_{i=1}^{p} f_i(X_i) + \varepsilon \tag{1}$$

where it is possible that $f_i(X_i) = 0$ for some $i \in \{1, \cdots, p\}$. Assuming Equation (1), additive model regression could be used to test if the response variable Y and some predictors $\boldsymbol{X}_a \subseteq \boldsymbol{X}$ are independent conditional on the other predictors $\boldsymbol{X}_b = \boldsymbol{X} \setminus \boldsymbol{X}_a$, because Y is independent of $\boldsymbol{X}_a$ given $\boldsymbol{X}_b$ if and only if $\mathrm{E}[Y|\boldsymbol{X}]$ is constant in $\boldsymbol{X}_a$}.

Additive regression works well as a conditional independence test in the study of Granger causality when no contemporaneous causation is allowed among time series, because the only type of conditional independence relations to be tested is the one described above. For example, in Bell et al. (1996), two additive models were fitted: one model for estimating the conditional expectation of a variable X_{T+1} given its T lags $\{X_1, X_2, \cdots, X_T\}$, another for conditional expectation of X_{T+1} given $\{X_1, X_2, \cdots, X_T\}$ and $\{Y_1, Y_2, \cdots, Y_T\}$. The F test was used to compare these two regression models: if the test failed to reject the first model, X_{T+1} was judged independent of $\{Y_1, Y_2, \cdots, Y_T\}$ given $\{X_1, X_2, \cdots, X_T\}$.

However, the use of additive model regression as a general purpose nonlinear conditional independence test is problematic, even for variables that are known to be related via additive models. Generally speaking, it is not always valid to use additive model regression to test conditional independence relations other than those between the response variable and some predictors given the other predictors. First, in some cases, additive model regression may miss some conditional dependencies. Consider a causal system with two exogenous variables X_1 and X_2, and an endogenous variable Y such that $Y = X_1^2 + X_2^2 + \varepsilon_Y$, where X_1, X_2 and ε_Y are independent Gaussian variables. Although the predictors X_1 and X_2 are dependent given the response variable Y, the conditional expectation of X_1 given Y and X_2 estimated using additive model regression will be constant in X_2. Second, even worse, in some cases additive model regression may miss some conditional independencies. Consider a system with two exogenous variables X_1 and X_2, and five endogenous variables $W = X_1 + X_2 + \varepsilon_W$, $Y = W^2 + \varepsilon_Y$, $U = \log(X_1) + \varepsilon_U$, $V = \log(X_2) + \varepsilon_V$, and $Z = U + V + \varepsilon_Z$. Although the two response variables Y and Z are independent conditional on the predictors X_1 and X_2, Z will be present in the conditional expectation of Y given $\{X_1, X_2, Z\}$ estimated by additive model regression. (Note that Y contains a term $2X_1X_2$, and $e^Z = e^{\varepsilon_U + \varepsilon_V + \varepsilon_Z} X_1 X_2$.)

Nevertheless, additive model regression has some very attractive features. First, and probably most importantly, it is not subject to the curse of dimensionality. In fact, Stone (1985) shows that the rate of convergence for an additive model regression is the same as that for a univariate smoother, which is much faster than a general multidimensional nonparametric regression method. The second major advantage of additive model regression is that it is possible to identify the contribution of each predictor to the response variable, thus allowing an intuitive interpretation of the fitted models.

In the following sections, we define a additive non-linear time series model by imposing linear constraints only among contemporaneous variables. We show that two families of conditional independence relations can be tested consistently among variables in a additive non-linear time series model using additive model regression. That is, asymptotically, additive model regression will neither miss any conditional independence relations nor report any spurious conditional independence relations when applied to data generated from a additive non-linear time series model to test those two families of conditional independence relations. We propose an inference procedure for nonlinear time series data that requires only information about these two families of conditional independence relations.

2. Additive Non-linear Time Series Models

Below we present the definition of a family of nonlinear time series models for which additive model regression based conditional independence test is possible. Here $\boldsymbol{X}_t$ is a p dimensional observed time series, $\boldsymbol{U}_t$ a q dimensional unobserved time series, and $\boldsymbol{\varepsilon}_t$ a p dimensional white noise.
Definition: A p dimensional time series $\{\boldsymbol{X}\}_t = \{\cdots, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_T, \cdots\}$, where $\boldsymbol{X}_t = \{X_{t,1}, \cdots, X_{t,p}\}$, is generated from a *lag T additive non-linear model* if it satisfies the following conditions:

C1 For $i = 1, \cdots, p$,

$$X_{t,i} = \sum_{1 \le j \le p, j \ne i} c_{j,i} X_{t,j} + \sum_{1 \le k \le p, 1 \le l \le T} f_{k,i,l}(X_{t-l,k}) + \sum_{m=1}^{q} b_{m,i} U_{t,m} + \varepsilon_{t,i} \tag{2}$$

where $b_{m,i}$'s and $c_{j,i}$'s are constants, and $f_{k,i,l}$'s are smooth univariate functions

C2 $\cdots, \varepsilon_{1,1}, \cdots, \varepsilon_{1,p}, \varepsilon_{2,1}, \cdots, \varepsilon_{t,i}, \cdots$ and $\cdots, U_{1,1}, \cdots, U_{1,q}, U_{2,1}, \cdots, U_{t,j}, \cdots$ are jointly independent, with $\varepsilon_{t,i} \sim N(0, \sigma^2_{1,i})$ and $U_{t,j} \sim N(0, \sigma^2_{2,j})$.

C3 There is a k and an i such that $f_{k,i,T}(\cdot) \ne 0$

C4 There is no sequence of indices $\{j_1, j_2, \cdots, j_m\}$ such that c_{j_1,j_2}, c_{j_2,j_3}, $\cdots$, c_{j_{m-1},j_m}, c_{j_m,j_1} are all nonzero.

The model is additive because Equation (2) includes both linear terms and arbitrary smooth terms. It is also recursive in the sense that given an initialization of $\boldsymbol{X}_{t-T}, \cdots, \boldsymbol{X}_{t-1}$, all the later points in the time series, starting from $\boldsymbol{X}_t$, can be generated inductively.

A additive non-linear model can be causally interpreted in the following way:

- $X_{t,j}$ is a direct cause of $X_{t,i}$ if and only if $c_{j,i} \ne 0$ in Equation (2), (for the definition of *direct cause*, see Spirtes et al., 2000; Pearl, 2000);
- $X_{t-l,j}$ is a direct cause of $X_{t,i}$ if and only if $f_{j,i,l}(\cdot) \ne 0$ in Equation (2);
- Latent common causes are allowed only for variables in the same time tier, and $X_{t,i}$ and $X_{t,j}$ have a latent common cause $U_{t,m}$ if and only if there is an m such that $b_{m,i} b_{m,j} \ne 0$.

  Note that both $\boldsymbol{U}_t$ and $\boldsymbol{\varepsilon}_t$ are multi-dimensional Gaussian white noise and both are unobserved. However, for $i = 1, \cdots, p$, $\varepsilon_{t,i}$ can only be a direct cause of $X_{t,i}$, where for $m = 1, \cdots, q$, $U_{t,m}$ can be a direct cause of several variables in $\boldsymbol{X}_t$.
- Condition C4 means that no contemporaneous feedback is allowed. If condition C4 is violated, X_{t,j_m} would be a direct cause of X_{t,j_1}, while at the same time X_{t,j_1} would be a (possibly indirect) cause of X_{t,j_m}.

  Note that using results of Richardson and Spirtes (1999) the method described in Section 3 can be modified to allow contemporaneous feedback.

A additive non-linear model can be represented by a directed graph consisting of nodes for $X_{T+1,1}, \cdots, X_{T+1,p}$ and their direct causes, and directed edges between nodes for the direct influences between the corresponding variables. We call this graph a *unit causal graph* for the corresponding

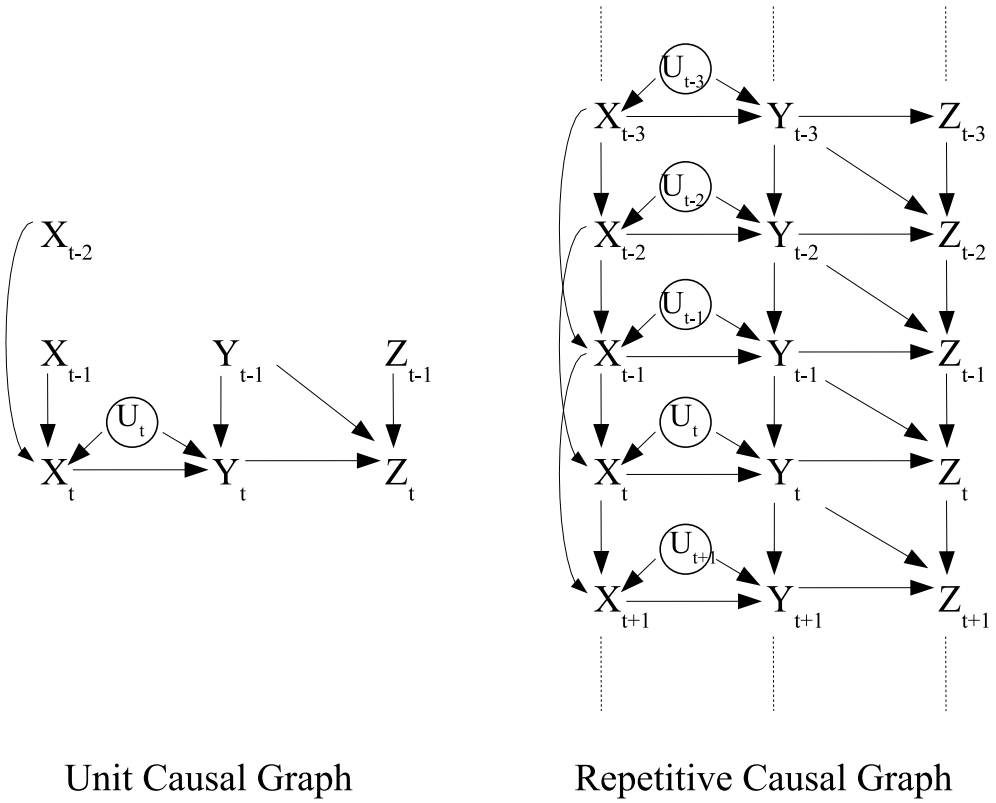

Figure 1: Unit causal graph and repetitive causal graph

time series. A unit causal graph can be extended to a *repetitive causal graph* by including all the variables in $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{T+1}$. Moreover, if there is an edge between $X_{T+1,i}$ and $X_{t,j}$, where $1 \leq t \leq T+1$, then similar edges will be added between $X_{T+1-l,j}$ and $X_{t-l,i}$ for $1 \leq l \leq t-1$. Figure 1 shows a unit causal graph and a segment of the corresponding repetitive graph. (The circled variables are latent variables.) In the remaining part of this paper, all time series causal models are represented by unit causal graphs.

Additive non-linear time series models make it possible to use the additive regression method, which is not subject to the curse of dimensionality, to test conditional independence for nonlinear time series. For a time series $\{\boldsymbol{X}\}_t$ generated from a lag T additive non-linear model, the following holds:

Proposition 1: Let X_t^1 and X_t^2 be any two distinct entries of random vector $\boldsymbol{X}_t$, $\boldsymbol{X}_t^c$ any subset, possibly empty, of $\boldsymbol{X}_t \setminus \{X_t^1, X_t^2\}$, and $\boldsymbol{X}_t^d$ any subset, possibly empty, of $\boldsymbol{X}_t \setminus \{X_t^1\}$. Let $\boldsymbol{X}^l = \{\boldsymbol{X}_{t-T}, \cdots, \boldsymbol{X}_{t-1}\}$, and $\boldsymbol{X}^e = \boldsymbol{X}^l \setminus \{X_{t-i,j}\}$ for some $X_{t-i,j} \in \boldsymbol{X}^l$.

- For any $\boldsymbol{x}_t^d$ and $\boldsymbol{x}^l$, conditional on $\boldsymbol{X}_t^d = \boldsymbol{x}_t^d$ and $\boldsymbol{X}^l = \boldsymbol{x}^l$, X_t^1 has a normal distribution $N(\mu_{1|a}, \sigma_{1|a}^2)$ such that $\mu_{1|a}$ is a linear combination of $\boldsymbol{x}_t^d$ and smooth univariate functions of entries of $\boldsymbol{x}^l$, and $\sigma_{1|a}$ is independent of t, $\boldsymbol{x}^l$ and $\boldsymbol{x}_t^d$. Thus, X_t^1 is independent of $X_{t-i,j}$ conditional on $\boldsymbol{X}^e$ and $\boldsymbol{X}_t^d$ if and only if $\mu_{1|a}$, the conditional expectation of X_t^1 given $\boldsymbol{X}_t^d = \boldsymbol{x}_t^d$ and $\boldsymbol{X}^l = \boldsymbol{x}^l$, is constant in $x_{l,j}$.

- For any x_t^2, $\boldsymbol{x}_t^c$ and $\boldsymbol{x}^l$, conditional on $X_t^2 = x_t^2$, $\boldsymbol{X}_t^c = \boldsymbol{x}_t^c$, and $\boldsymbol{X}^l = \boldsymbol{x}^l$, X_t^1 has a normal distribution $N(\mu_{1|b}, \sigma_{1|b}^2)$ such that $\mu_{1|b}$ is a linear combination of x_t^2, $\boldsymbol{x}_t^c$, and smooth univariate functions of entries of $\boldsymbol{x}^l$, and $\sigma_{1|b}$ is constant in t, x_t^2, $\boldsymbol{x}_t^c$, and $\boldsymbol{x}^l$. Thus, X_t^1 is independent of X_t^2 conditional on $\boldsymbol{X}_t^c$ and $\boldsymbol{X}^l$ if and only if, $\mu_{1|b}$, the conditional expectation of X_t^1 given $X_t^2 = x_t^2$, $\boldsymbol{X}_t^c = \boldsymbol{x}_t^c$, and $\boldsymbol{X}^l = \boldsymbol{x}^l$, is constant in x_t^2.

Proposition 1 implies that it is possible to use additive model regression to test the following two types of conditional independence relations among variables in a additive non-linear model. First, we can test if X_t^1 and X_t^2 are independent conditional on $\boldsymbol{X}_t^c$ and $\boldsymbol{X}^l$ by estimating the conditional expectation of X_t^1 given $\{X_t^2\} \cup \boldsymbol{X}_t^c \cup \boldsymbol{X}^l$ using additive model regression, and check if X_t^2 is a significant predictor for X_t^1 using statistical tests such as the F test (Bell et al., 1996) or the BIC scores (Huang and Yang, 2004). Similarly, if $X_{t-i,j}$ is not a significant predictor for X_t^1 in the additive model regression of X_t^1 against $\boldsymbol{X}^l$ and $\boldsymbol{X}_t^d$, we would say X_t^1 and $X_{t-i,j}$ are independent conditional on $\boldsymbol{X}_t^d$ and $\boldsymbol{X}^e$.

To make the above tests valid, we also need the assumption that additive model regression is an (asymptotically) consistent estimator of conditional expectations such as $\mathrm{E}[X_t^1|\boldsymbol{X}_t^d, \boldsymbol{X}^l]$ and $\mathrm{E}[X_t^1|X_t^2, \boldsymbol{X}_t^c, \boldsymbol{X}^l]$. Fortunately, it has been shown that, given a stationary nonlinear time series $\{\boldsymbol{X}\}_t$, nonparametric estimation of the conditional mean $\mathrm{E}[X_t|\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-T}]$ is asymptotically consistent and/or asymptotically normal, provided certain conditions are satisfied (Robinson, 1983; Truong and Stone, 1992; Chen and Tsay, 1993; Tjøstheim and Auestad, 1994; Härdle et al., 1997; Cai and Masry, 2000; Huang and Yang, 2004). Generally speaking, besides some regularity conditions on the density of $\boldsymbol{X}_t \cup \boldsymbol{X}^l$ and smoothness condition on $\mathrm{E}[X_t|\boldsymbol{X}^l]$, $\{\boldsymbol{X}\}_t$ should satisfy some form of α mixing condition. $\{\boldsymbol{X}\}_t$ is α mixing if for some $\alpha(n) \to 0$,

$$\sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_t, B \in \mathcal{G}_{n+t}\} \leq \alpha(n)$$

where $\mathcal{F}_t$ is the σ-field generated by $\boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \cdots$, and $\mathcal{G}_{n+t}$ the σ-field generated by $\boldsymbol{X}_{t+n}, \boldsymbol{X}_{t+n+1}, \cdots$.

A concept closely related to α mixing is geometric ergodicity. A stationary time series $\{\boldsymbol{X}\}_t$ is geometrically ergodic if there is a function $M(x) < \infty$ and a constant $\rho < 1$ such that for all x:

$$\sup_A |P(X_n \in A|X_0 = x) - \pi(A)| \leq M(x)\rho^n$$

where π is the stationary distribution of $\{\boldsymbol{X}\}_t$. For stationary time series, geometric ergodicity implies α mixing for an $\alpha(n)$ of exponential rate (Davydov, 1973). Sufficient conditions for a nonlinear time series to be geometrically ergodic can be found in Chan and Tong (1994), An and Huang (1996), and Cline and Pu (1999). In particular, Xia and An (1999) provides a set of sufficient conditions for the geometric ergodicity of time series generated by projection pursuit models, of which our additive non-linear model is a special case.

3. A Causal Inference Algorithm

Consider a time series $\{\boldsymbol{X}\}_t = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_t, \cdots\}$ are generated from a lag T additive non-linear model. Let $\boldsymbol{X}^l = \{\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-T}\}$, X_t^1 and X_t^2 be any two entries of $\boldsymbol{X}_t$, $\boldsymbol{X}_t^b$ be any subset, possibly empty, of $\boldsymbol{X}_t \setminus \{X_t^1\}$, $\boldsymbol{X}_t^c$ be any subset, possibly empty, of $\boldsymbol{X}_t \setminus \{X_t^1, X_t^2\}$, $X_{t-i,j}$ any variable in $\boldsymbol{X}^l$, and $\boldsymbol{X}^e = \boldsymbol{X}^l \setminus \{X_{t-i,j}\}$. Using additive model regression, we can test two types of conditional independence relations: 1), if X_t^1 and X_t^2 are independent given $\boldsymbol{X}_t^c$ and $\boldsymbol{X}^l$, and 2), if X_t^1 and $X_{t-i,j}$ are independent given $\boldsymbol{X}_t^b$ and $\boldsymbol{X}^e$. These pieces of information are not generally sufficient for currently available causal inference algorithms, such as the PC and FCI, to be informative: these procedures require (in the worst case) complete conditional independence information. However, starting from the same principle behind the PC and FCI algorithms, we describe a procedure that requires only these two types of conditional independence information. The procedure, which is capable of producing very informative causal structures, takes advantage of the constraints on possible causal relations among the random variables imposed by additive non-linear models, for example, $X_{t_2,k}$ cannot be a cause of $X_{t_1,j}$ if $t_1 < t_2$, no latent common cause exists for $X_{t_2,k}$ and $X_{t_1,j}$ if $t_1 \neq t_2$, etc.

The following propositions are needed to justify our procedure. We assume familiarity with notions from the graphical modeling literature, including the notion of d-separation (Pearl, 2000), and faithfulness (Spirtes et al., 2000). In summary:

Formally a causal graph G is defined as an ordered pair $\langle \boldsymbol{V}, \boldsymbol{E} \rangle$, where $\boldsymbol{V}$ is the set of variables in G, and $\boldsymbol{E}$ the set of edges in G. An edge e in $\boldsymbol{E}$ is again defined as an ordered pair $\langle V_i, V_j \rangle$, where V_i and V_j are two variables in $\boldsymbol{V}$. Given an edge $e = \langle V_i, V_j \rangle$ in graph G, we say that V_i is a direct cause of V_j in G. The *subgraph G_m induced by $\boldsymbol{V}_m$*, where $\boldsymbol{V}_m$ is a subset of $\boldsymbol{V}$, is defined as an ordered pair $\langle \boldsymbol{V}_m, \boldsymbol{E}_n \rangle$ such that an edge $e = \langle V_i, V_j \rangle$ is in $\boldsymbol{E}_n$ if and only if e is in $\boldsymbol{E}$ and the two variables $\{V_i, V_j\}$ are both in $\boldsymbol{V}_m$. A vertex is a collider on an undirected path in a directed acyclic graph (DAG) if and only if it is the second member of both of two edges on the path, that is, two edges on the path are directed into it. Two vertices X, Y (representing random variables) are d-separated with respect to a set $\boldsymbol{Z}$ of vertices if and only if every undirected path between the variables contains a collider having no directed path into a member of $\boldsymbol{Z}$ or contains a non-collider that is a member of $\boldsymbol{Z}$. A joint distribution on the variables (vertices) of a DAG is faithful if and only if all conditional independence relations follow from the d-separation property applied to the DAG.

In the three propositions below, $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_t, \cdots\}$ form a time series generated from a lag T additive non-linear model, $\boldsymbol{X}^l = \{\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-T}\}$, X_t^1 and X_t^2 are any two entries of $\boldsymbol{X}_t$, and $\boldsymbol{X}^e = \boldsymbol{X}^l \setminus \{X_{t-i,j}\}$ for some $X_{t-i,j} \in \boldsymbol{X}^l$

Proposition 2: The d-separation relations among the variables in $\boldsymbol{X}_t$ conditional on $\boldsymbol{X}^l$ in a repetitive causal graph G_c are the same as the d-separation relations among the variables in $\boldsymbol{X}_t$ in the subgraph of G_c induced by $\boldsymbol{X}_t$.

Proof: See Moneta (2003), proposition 4. □

Proposition 3: Consider a time series $\{\boldsymbol{X}\}_t = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_t, \cdots\}$ generated from a lag T additive non-linear model. Let $\boldsymbol{X}^l = \{\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-T}\}$, X_t^1 and X_t^2 be any two entries of $\boldsymbol{X}_t$. Assuming faithfulness, if there is a variable $X_{t-i,j} \in \boldsymbol{X}^l$ such that X_t^2 and $X_{t-i,j}$ are independent conditional on $\boldsymbol{X}^e = \boldsymbol{X}^l \setminus \{X_{t-i,j}\}$, but $X_{t-i,j}$ and X_t^1 are not independent conditional on $\boldsymbol{X}^e$, then X_t^1 is not a cause of X_t^2.

Proof: Suppose X_t^1 is a cause of X_t^2, then there must be a directed path P' from X_t^1 to X_t^2 such that each vertex on P' is in $\boldsymbol{X}_t$. If $X_{t-i,j}$ and X_t^1 are dependent given $\boldsymbol{X}^e$, there must be a path P d-connecting $X_{t-i,j}$ and X_t^1 given $\boldsymbol{X}^e$. Thus, no variable in $\boldsymbol{X}^e$ is a non-collider on path P, and all the colliders on path P must be observed ancestors of $\boldsymbol{X}^e$, hence must be in $\boldsymbol{X}^e$. (Note that the set of observed ancestors of $\boldsymbol{X}^e$ is either $\boldsymbol{X}^e$ or $\boldsymbol{X}^e \cup \{X_{t-i,j}\}$). This implies that P must be into X_t^1, because otherwise either P would be a direct path from X_t^1 to $X_{t-i,j}$, which is not allowed, or there must be a collider on P that is both a descendant of X_t^1 and an element of $\boldsymbol{X}^e$, which also is impossible. By appending the direct path P' to P, we get a path d-connecting $X_{t-i,j}$ and X_t^2 given $\boldsymbol{X}^e$, which is a contradiction. □

Proposition 4: Consider a time series $\{\boldsymbol{X}\}_t = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_t, \cdots\}$ generated from a lag T additive non-linear model. Let $\boldsymbol{X}^l = \{\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-T}\}$, X_t^1 be any entry of $\boldsymbol{X}_t$, $X_{t-i,j}$ be any variable in $\boldsymbol{X}^l$, $\boldsymbol{X}_t^d$ be the set of all observed contemporary direct causes of X_t^1, and $\boldsymbol{X}^e = \boldsymbol{X}^l \setminus \{X_{t-i,j}\}$. Assuming faithfulness, $X_{t-i,j}$ and X_t^1 are dependent conditional on $\boldsymbol{X}_t^d$ and $\boldsymbol{X}^e$ if and only if:

- either $X_{t-i,j}$ is a direct cause of X_t^1,
- or there is a path P between X_t^1 and $X_{t-i,j}$, with $\langle W_1, \cdots, W_m \rangle$ being the set of observed variables on P between X_t^1 and $X_{t-i,j}$ and ordered along the direction from X_t^1 to $X_{t-i,j}$, such that:
  1. $W_i \in \boldsymbol{X}_t$ for $i = 1, \cdots, m$;
  2. X_t^1 and W_1 have a latent common cause;
  3. if $W_i \in \boldsymbol{X}_t^d$ then W_i is a collider on P;
  4. W_i is a (possibly indirect) cause of X_t^1 for $i = 1, \cdots, m$;
  5. $X_{t-i,j}$ is a direct cause of W_m.

Proof: The *if* part of the proposition is trivial, here we only prove the *only if* part.

Suppose $X_{t-i,j}$ is not a direct cause of X_t^1, then there is a path P d-connecting $X_{t-i,j}$ and X_t^1 conditional on $\boldsymbol{X}_t^d$ and $\boldsymbol{X}^e$. Let $\boldsymbol{W} = \langle W_1, \cdots, W_m \rangle$ be the set of observed variables on P between X_t^1 and $X_{t-i,j}$, ordered along the direction from X_t^1 to $X_{t-i,j}$.

To show that $W_i \in \boldsymbol{X}_t$ for $i = 1, \cdots, m$, we note that if W_j is the first element in $\boldsymbol{W}$ such that $W_j \notin \boldsymbol{X}_t$, it must belong to $\boldsymbol{X}^e$, where W_{j-1} is in $\boldsymbol{X}_t$. Because there is no observed variable between W_{j-1} and W_j on P, by the definition of additive non-linear models, there must be a direct edge from W_j to W_{j-1} on P (let $X_t^1 = W_0$ when $j = 1$). This means that W_j is not a collider on P, hence P cannot d-connect X_t^1 and $X_{t-i,j}$ conditional on $\boldsymbol{X}^e$ and $\boldsymbol{X}_t^d$, which contradicts our assumption. Using the same argument, given that $W_m \in \boldsymbol{X}_t$, it is easy to see that $X_{t-i,j}$ must be a direct cause of W_m.

Next we show that W_1 and X_t^1 must have a latent common cause. Assume that there is no latent common cause for W_1 and X_t^1. Because there is no observed variable between W_1 and X_t^1 on P, they must be adjacent on P, hence there must be a direct causal relation between X_t^1 and W_1. Consider the two alternative cases:

- First, suppose that W_1 is a direct cause of X_t^1. Then $W_1 \in \boldsymbol{X}_t^d$, and is a non-collider on P, hence P cannot d-connecting $X_{t-i,j}$ and X_t^1 conditional on $\boldsymbol{X}_t^d$ and $\boldsymbol{X}^e$.

- Second, suppose X_t^1 is a direct cause of W_1. Then there must be a variable W_i for some $i \geq 1$ such that the subpath $\{X_t^1, W_1, \cdots, W_i\}$ of P is a directed path from X_t^1 to W_i, and W_i is a collider on P. This would imply that W_i has to be a cause of X_t^1, for otherwise neither W_i nor any of its descendants belong to $\boldsymbol{X}_t^d$, which means that P cannot d-connect X_t^1 and $X_{t-i,j}$ conditional on $\boldsymbol{X}^e$ and $\boldsymbol{X}_t^d$. But allowing W_i to be a cause of X_t^1 would make the path $X_t^1, W_1, \cdots, W_i, X_t^1$ a directed cycle, which is impossible.

It is obvious that if $W_i \in \boldsymbol{X}_t^d$, then it must be a collider on P. To show that W_i is a cause of X_t^1, we note that if W_j is a collider on P, it must be a cause of X_t^1, for otherwise neither W_j nor any of its descendants belongs to $\boldsymbol{X}_t^d$, hence P cannot d-connect X_t^1 and $X_{t-i,j}$ conditional on $\boldsymbol{X}^e$ and $\boldsymbol{X}_t^d$. Therefore W_i must be a cause of X_t^1, because it is either a collider on P, or a cause of a collider on P. □

Given propositions 2, 3, and 4, we propose a three-step procedure for inference to unit causal graphs from time series data generated by additive non-linear models. The output of this causal inference procedure is a Partial Ancestral Graph (PAG). Roughly speaking, a PAG is a graph consisting of a list of vertices representing observed random variables, and 3 types of end points, $-$, $\circ$, and $>$, which are combined to form the following 4 types of edges representing causal relations between random variables.

- $X \to Y$ means that X is a (possibly indirect) cause of Y.
- $X \leftrightarrow Y$ means that there is a latent variable Z that is a (possibly indirect) cause of both X and Y.
- $X \circ\!\!\to Y$ means either $X \to Y$ or $X \leftrightarrow Y$.
- $X \circ\!\!-\!\!\circ Y$ means either $X \to Y$, or $Y \circ\!\!\to X$. In other words, $X \circ\!\!-\!\!\circ Y$ means that X and Y cannot be d-separated by any other observed variables.

For detailed explanation of PAGs, see Spirtes et al. (2000). Following Spirtes et al. (2000), we also use * as a meta symbol to represent any of the three end points.

Below is a constraint based additive non-linear time series causal inference procedure for non-linear time series with latent common causes. The conditional independence information required by the procedure can be obtained using additive model regression based conditional independence tests mentioned in the previous section. Here we assume that the time series data satisfies various conditions for the asymptotic consistency and normality of the additive model estimator, and that an upper bound T_{max} on the unknown true lag number T of the additive non-linear model has been set, either using the procedures in Tjøstheim and Auestad (1994) or Huang and Yang (2004), or based on background knowledge. So long as T_{max} is no less than T, the following procedure asymptotically obtains a correct PAG. Of course, choosing a T_{max} much higher than T will reduce the efficiency of the procedure.

The symbols in the following procedure are defined in the same way as in the beginning of this section, except that $\boldsymbol{X}^l$ is redefined as $\boldsymbol{X}^l = \{\boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-T_{max}}\}$.

1. Identify contemporary causal relations

   (a) For all choices of X_t^1, X_t^2, and $\boldsymbol{X}_t^c$, determine if X_t^1 is independent of X_t^2 conditional on $\boldsymbol{X}_t^c$ and $\boldsymbol{X}^l$.

(b) Treat the above conditional independence relations as if they were conditional independence relations between X_t^1 and X_t^2 given $\boldsymbol{X}_t^c$.

- Feed these conditional independencies to a causal inference algorithm allowing presence of latent common causes, such as the FCI algorithm. Derive the PAG for the contemporary causal structure among variables in $\boldsymbol{X}_t$. Call this PAG π_t.
- For all choices of X_t^1, identify the set of *possible contemporaneous direct causes* of X_t^1, where X_t^2 is a possible contemporaneous direct cause of X_t^1 if in π_t either $X_t^2 \circ\!\!-\!\!\circ X_t^1$, or $X_t^2 \circ\!\!\rightarrow X_t^1$, or $X_t^2 \rightarrow X_t^1$. Denote by $\mathrm{PCDC}(X_t^1)$ the set of possible contemporaneous direct causes of X_t^1.

2. Identify lagged causal relations.

(a) Create a new graph π_f such that the vertices in π_f are $\{\boldsymbol{X}_t, \boldsymbol{X}_{t-1}, \cdots, \boldsymbol{X}_{t-T}\}$, and the edges in π_f are exactly the same as the edges in π_t.

(b) For all choices of X_t^1, $X_{t-i,j}$, and $\boldsymbol{X}_t^b$, determine if X_t^1 and $X_{t-i,j}$ are independent given $\boldsymbol{X}^e$ and $\boldsymbol{X}_t^b$

- For all choices of X_t^1, identify the set of *possible lagged direct causes* of X_t^1, where a lagged variable $X_{t-i,j}$ is a possible lagged direct cause of X_t^1 if for all $\boldsymbol{X}_t^d \subseteq \mathrm{PCDC}(X_t^1)$, $X_{t-i,j}$ and X_t^1 are dependent given $\boldsymbol{X}_t^d$ and $\boldsymbol{X}^e$. Denote by $\mathrm{PLDC}(X_t^1)$ the set of possible lagged direct causes of X_t^1
- For all choices of X_t^1, identify the set of *permanent lagged predictors* of X_t^1, where $X_{t-i,j}$ is a permanent lagged predictor of X_t^1 if for all $\boldsymbol{X}_t^b \subseteq (\boldsymbol{X}_t \setminus \{X_t^1\})$, $X_{t-i,j}$ and X_t^1 are dependent given $\boldsymbol{X}_t^b$ and $\boldsymbol{X}^e$. Denote by $\mathrm{PLP}(X_t^1)$ the set of permanent lagged predictors of X_t^1

(c) Add edges representing the lagged causes of each variable in $\boldsymbol{X}_t$ to π_f:

i. For all choices of X_t^1, add an edge $X_{t-i,j} \rightarrow X_t^1$ to π_f if $X_{t-i,j} \in \mathrm{PLP}(X_t^1)$.

ii. For all choices of X_t^1, add an edge $X_{t-i,j} \rightarrow X_t^1$ to π_f if $X_{t-i,j} \in \mathrm{PLDC}(X_t^1)$, and $X_{t-i,j}$ is not adjacent to any other variable in π_f.

3. Orient the contemporary PAG according to the following rule:

(a) Repeat the following procedure until no more changes can be made to π_f.

i. If $X_{t-i,j} \rightarrow X_t^1 \circ\!\!-\!\!* X_t^2$ is in π_f, and $X_{t-i,j}$ and X_t^2 are not adjacent, then:
If $X_{t-i,j}$ and X_t^2 are independent given $\boldsymbol{X}^e$, but dependent given X_t^1 and $\boldsymbol{X}^e$, then orient the edge between X_t^1 and X_t^2 as $X_t^1 \leftarrow\!\!* X_t^2$

ii. If $X_{t-i,j} \rightarrow X_t^1 \circ\!\!-\!\!* X_t^2$ is in π_f, and $X_{t-i,j}$ and X_t^2 are not adjacent, then:
If $X_{t-i,j}$ and X_t^2 are dependent conditional on $\boldsymbol{X}^e$, but independent conditional on X_t^1 and $\boldsymbol{X}^e$, then orient the edge between X_t^1 and X_t^2 as $X_t^1 \rightarrow X_t^2$

(b) Apply the orientation step of FCI algorithm to further orient the contemporary PAG π_f.

Proposition 2 provides justification for the first step in this procedure, proposition 3 the third step. Proposition 4 is needed for the second step, as we can see that the set of contemporaneous direct causes of a variable X_t^1 is a subset of $\mathrm{PCDC}(X_t^1)$, thus by proposition 4 we have:

$$\text{Lagged direct causes of } X_t^1 \subseteq \text{PLP}(X_t^1) \subseteq \text{PLDC}(X_t^1) \subseteq \text{Lagged causes of } X_t^1$$

Note that step 2(c) is designed to make the procedure more robust.

The complexity of the above procedure is primarily determined by step 1(a), where $k2^{k-1}$ additive model regressions are performed to test the conditional independence relations required by the later steps.

We want to emphasize that the above procedure can be modified in various ways to accommodate changes in the assumptions about the time series data generating models. In the last section (Section 6) of this paper, we discuss in details about different extensions of the above procedure.

4. Simulation Study

In this section, we conduct a simple simulation study to evaluate the performance of the additive non-linear causal inference algorithm presented in Section 3. In particular, we would like to see if the additive non-linear algorithm can provide a viable solution to the problem of nonlinear time series causal inference. For comparison, we also apply a causal inference procedure designed for linear time series to the simulated data. Because there is no currently available efficient *automated* causal inference algorithm for linear time series with contemporaneous causal relations, the linear procedure used for comparison actually is an extension of our additive non-linear causal inference procedure under the assumption that the time series data are generated from linear models. (Bessler et al. 2002, Demiralp and Hoover 2003, Moneta 2003 and Hoover 2005 discussed efficient ways of identifying the contemporaneous causal pattern, that is, the Markov equivalence classes (MEC) of the causal graphs for contemporaneous variables assuming causal sufficiency. However, their procedures are not complete because, when the MEC consists of multiple contemporaneous causal graphs, these procedures all require further background information to uniquely identify the contemporaneous causal graph before proceeding to derive the causal pattern for both contemporaneous and lagged variables. Oxley et al. (2004) provides a less efficient algorithm for linear time series that treats a k-dimensional lag p structural vector autoregressive model (SVAR(p)) as a linear causal model with $k(p+1)$ variables.) The linear procedure differs from the additive non-linear algorithm only in step 1: unlike the original algorithm which uses additive regression to test conditional independence, the linear procedure uses linear regression instead.

We use the Mersenne Twister algorithm implemented in java package RngPack (version 1.1a) for random number generation, and the `gam` function in the R package gam (version 0.97) for additive model regression.

The simulated data are generated from the four causal structures shown in Figure 2. Note that in this simulation study the true PAGs happen to have no circles, and can be represented by the same graphs in Figure 2. The chain-like contemporaneous causal structure is chosen to evaluate the ability of our algorithm to identify the direction of those contemporaneous causal relations that could not be detected using previous algorithms (Bessler et al., 2002; Demiralp and Hoover, 2003; Moneta, 2003; Hoover, 2005). For each causal structure, we consider the following four types of models, characterized by the type of functional relations between an effect variable and its direct causes:

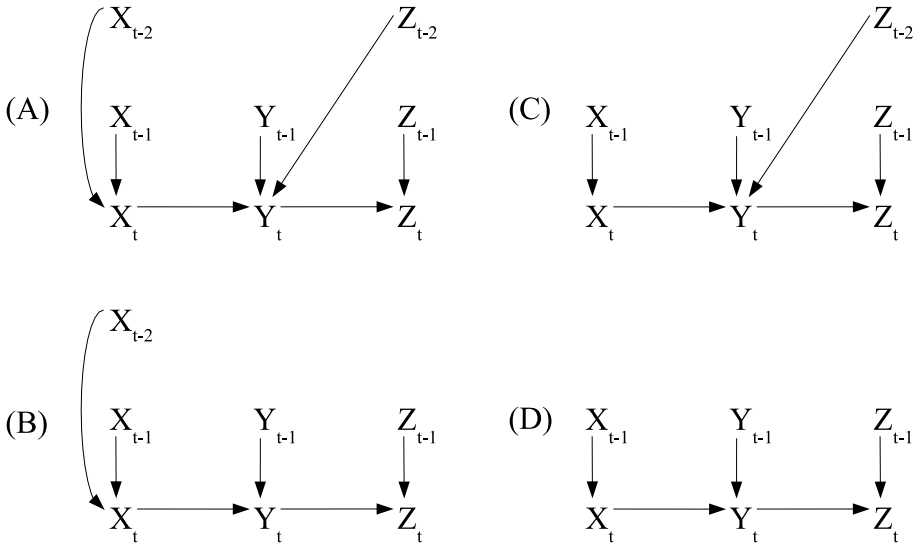

Figure 2: Causal graphs and true PAGs of simulation data

- Trigonometric lag models: Each contemporaneous variable is a linear combination of other contemporaneous variables and univariate trigonometric functions of lagged variables. For example, in one model, we have:

$$Y_t = 0.5X_t + \sin(2Y_{t-1}) - \cos(10Z_{t-2}) + \varepsilon_Y.$$

- Polynomial lag models: Each contemporaneous variable is a linear combination of other contemporaneous variables and univariate polynomial functions of lagged variables. For example, in one model, we have:

$$Y_t = 0.5X_t + 0.3Y_{t-1}^2 - 0.1Z_{t-2}^3 + \varepsilon_Y.$$

- Linear lag models: Each contemporaneous variable is a linear combination of other contemporaneous variables and lagged variables. For example, in one model, we have:

$$Y_t = 0.5X_t + 0.3Y_{t-1} - 0.1Z_{t-2} + \varepsilon_Y.$$

- Trigonometric contemporaneous models: Each contemporaneous variable is a linear combination of univariate trigonometric functions of other contemporaneous variables and lagged variables. For example, in one model, we have:

$$Y_t = \cos(X_t) + \sin(2Y_{t-1}) - \cos(10Z_{t-2}) + \varepsilon_Y.$$

Note that these models do not belong to the family of additive non-linear time series models, for they violate the assumption C1.

In total we have 16 data generating models, with 12 of them being additive non-linear time series models (including 4 linear time series models). For each of the 16 models, we generate 4 random time series data sets of length 200, 500, 1000, and 2000 respectively. For each data set, we run both the additive non-linear procedure and the linear procedure. The upper bound T_{max} of the true lag number T is set to 3 for all simulations, (T is equal to 2 for 12 of the data generating models based on casual structure (A), (B), and (C) in Figure 2, and 1 for the other 4 models based on casual structure (D)). The learned PAGs are compared with the true PAGs, which are also represented by the graphs in Figure 2.

The additive non-linear procedure presented in Section 3 requires, for each contemporaneous variable, say X_t, the following two types of conditional independence information: (1) if X_t is independent of another contemporaneous variable, say Y_t, given all the lagged variables $\boldsymbol{L} = \{X_{t-2}, X_{t-1}, Y_{t-2}, Y_{t-1}, Z_{t-2}, Z_{t-1}\}$ and a subset of the remaining contemporaneous variables, say, $\{Z_t\}$; and (2), if X_t is independent of a lagged variable, say X_{t-1}, given all the other lagged variables and a subset of contemporaneous variables, say, $\{Z_t\}$. These conditional independence relations are tested by checking if $\mathrm{E}[X_t|\boldsymbol{L}, Z_t]$ is constant in Y_t or X_{t-1} respectively. For example, to test if X_{t-1} is present in $\mathrm{E}[X_t|\boldsymbol{L}, Z_t]$, we follow Huang and Yang (2004) by starting from a model A, where X_t is regressed against $\boldsymbol{L}$ and Z_t, and searching for a submodel of A with the lowest BIC score. If X_{t-1} is present in this submodel with lowest BIC score, it is present in $\mathrm{E}[X_t|\boldsymbol{L}, Z_t]$. Otherwise, it is not.

The simulation results are summarized in Figure 3. Each of the four panes in Figure 3 summarizes the results of 16 simulated time series data sets generated from the same type of models. We use the average error rates to evaluate the performance of the two algorithms. The definitions of the various error rates are similar to those in Spirtes and Meek (1995). Consider a p dimensional time series data. An edge omission error occurs when two variables are adjacent in the true PAG but not in the learned PAG. An edge commission error occurs when two variables are adjacent in the learned PAG but absent in the true PAG.

The edge omission error rate is defined as:

$$E_o = \frac{\text{Number of edge omission errors}}{\text{Number of edges in the true PAG}}.$$

The edge commission error rate is defined as:

$$E_c = \frac{\text{Number of edge commission errors}}{\text{Maximum number of possible edge commission errors}}.$$

When inferring causal structure from a p dimensional time series data set, if the upper bound of the true lag number is set to T_{max}, the maximum number of possible edge commission errors is equal to:

$$p^2 T_{max} + \frac{p(p-1)}{2} - \text{Number of edges in the true PAG}$$

where $p^2 T_{max} + p(p-1)/2$ is the maximum number of edges can be found in the unit causal graph for any p-dimension lag T_{max} time series model.

The solid lines in each pane of Figure 3 represent the average omission error rates for different time series lengths; the dotted lines represent the average commission error rates. Blue lines with

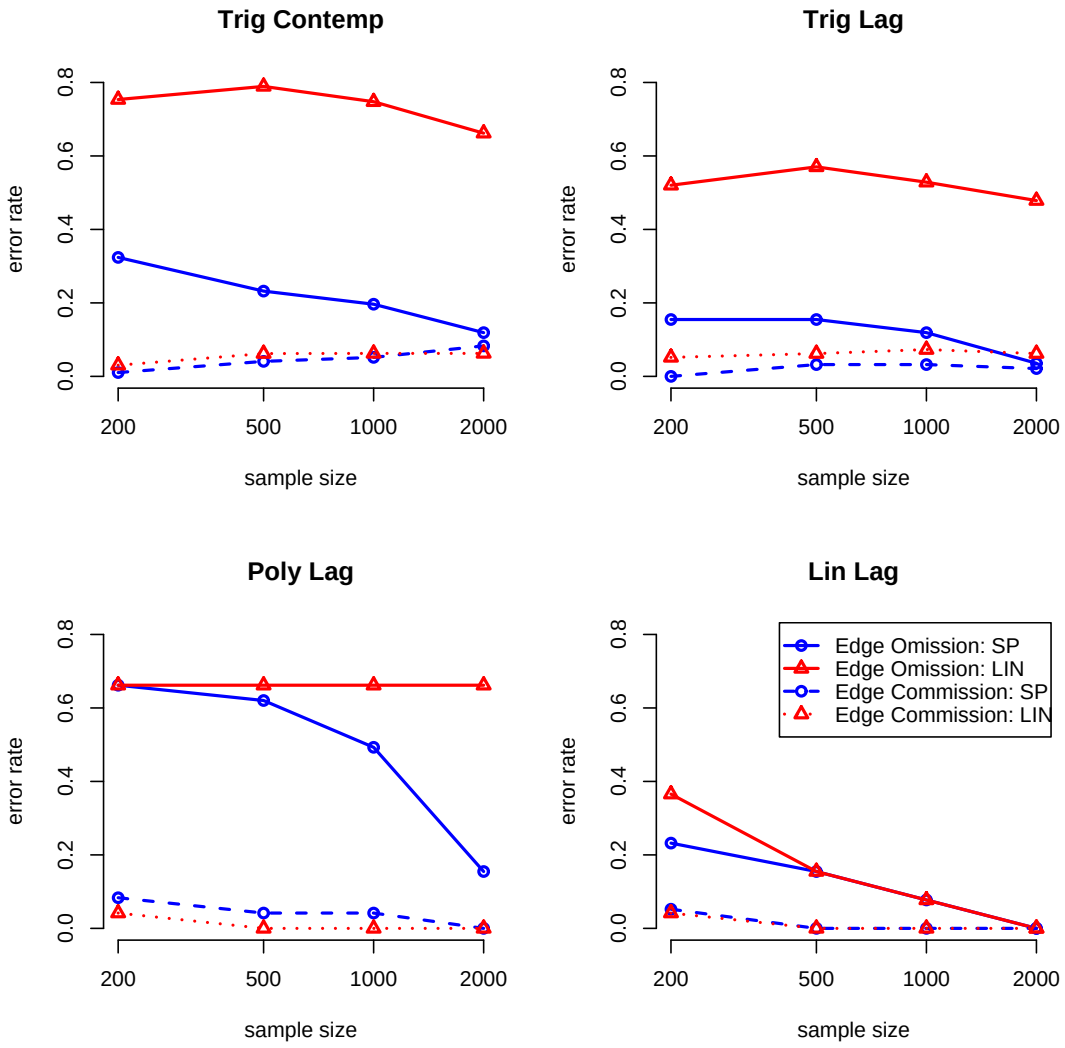

Figure 3: Error rate for edge discovery

circles represent results obtained by the additive non-linear algorithm, red lines with triangles the results by the linear procedure.

The pane with label "Trig Contemp" gives the results for data generated from the *trigonometric contemporaneous models*. We choose these models in the simulation study precisely because they lie outside of the family of additive non-linear time series models, for they violate the functional assumption (C1) in the definition of additive non-linear time series models. The simulation results suggest that, when the assumption C1 is violated, the additive non-linear algorithm can still discover most of the edges. However, as the length of time series increases, the average number of extra edges also increases, apparently because the data generating models are not additive non-linear time series models. The linear procedure is not satisfactory, missing most of the edges in the true models.

The panes labeled with "Trig Lag" and "Poly Lag" show the results for *trigonometric lag models* and *polynomial lag models*, both of which are genuine additive non-linear time series models. The additive non-linear algorithm performs very well for the trigonometric lag models, but less than satisfactory for polynomial lag models. Its performance for polynomial lag models, however, does improve as the length of time series increases. The linear procedure performs poorly in both cases, missing at least half of the edges.

The pane with label "Lin Lag" provides the results for *linear lag models*. Given that a linear lag model is simply a linear time series model, which is a special case of additive non-linear time series model, we expect that both algorithms should perform very well, as they do. This, on the one hand, suggests that the linear procedure is a good choice for linear time series causal inference, on the other hand, implies that the additive non-linear algorithm does not suffer from overfitting.

We also compare the average error rates for orientation of the edges among contemporaneous variables by the additive non-linear algorithm and the linear procedure. Suppose X_t and Y_t are adjacent in both the learned PAG and the true PAG, an arrowhead omission error occurs if the edge is oriented as $X_t *\!\!\rightarrow Y_t$ in the true PAG, but as $X_t *\!\!-\!\!* Y_t$ or $X_t *\!\!-\!\!\circ Y_t$ in the learned PAG. Similarly, an arrowhead commission error occurs if the edge is oriented as $X_t -\!\!* Y_t$ or $X_t \circ\!\!-\!\!* Y_t$ in the true PAG, but as $X_t \leftarrow\!\!* Y_t$ in the learned PAG. Let E be the set of edges among contemporaneous variables in the true PAG such that the pairs of variables connected by these edges are also adjacent in the learned PAG. The arrowhead omission error rate is defined as:

$$A_o = \frac{\text{Number of arrowhead omission errors}}{\sum_{e \in E} \text{Number of arrowheads in } e}.$$

The arrowhead commission error rate is defined as:

$$A_c = \frac{\text{Number of arrowhead commission errors}}{\sum_{e \in E} \text{Number of non-arrowheads in } e}.$$

In Figure 4, the solid lines in each pane represent the average *arrowhead omission error rates*; the dotted lines represent the average *arrowhead commission error rates*. As in Figure 3, blue lines with circles represent results obtained by the additive non-linear algorithm, red lines with triangles the results by the linear procedure. (Note that in the top two panels labeled respectively with "Trig Contemp" and "Trig Lag", the lines representing omission error and commission error for the additive non-linear algorithm overlap. In the bottom two panels labeled respectively with "Poly Lag" and "Lin Lag", the lines representing commission error for the additive non-linear algorithm and the linear algorithm overlap.)

There are two more scores to measure how close a learned PAG is to the true PAG, that is, the tail omission error rate and the tail commission error rate. Suppose X_t and Y_t are adjacent in both the learned PAG and the true PAG, a tail omission error occurs if the edge is oriented as $X_t \rightarrow Y_t$ in the true PAG, but as $X_t \circ\!\!-\!\!* Y_t$ in the learned PAG. A tail commission error occurs if the edge is oriented as $X_t \circ\!\!-\!\!* Y_t$ in the true PAG, but as $X_t \rightarrow Y_t$ in the learned PAG. Note that these definitions are stated so that an arrowhead commission/omission error will not be counted again as a tail omission/commission error. Because there is no circle in the true PAGs in this simulation study, we can only compute the tail omission errors for the learn PAGs, shown in Figure 5. The tail omission error rate is defined as:

$$T_o = \frac{\text{Number of tail omission errors}}{\sum_{e \in E} \text{Number of tails in } e}.$$

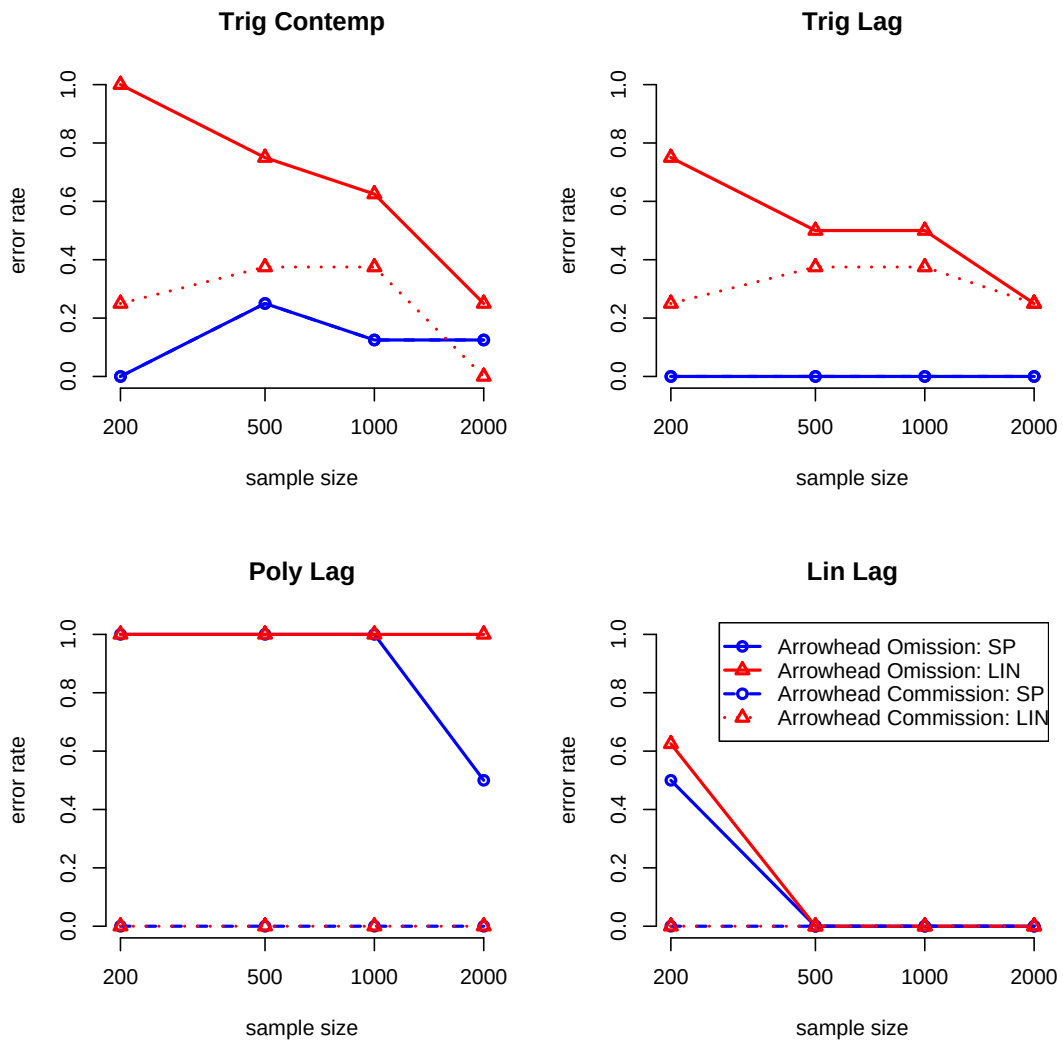

Figure 4: Error rate for edge orientation: Arrowhead

The additive non-linear algorithm gives excellent results. For example, for the data sets generated from additive non-linear models, that is, the trigonometric lag models, the polynomial lag models, and linear lag models, the additive non-linear algorithm makes no arrowhead commission errors. The linear procedure performs quite well for polynomial lag models and linear lag models.

Although the scope of this simulation study is very limited, we can get some general idea about the performance of our additive non-linear casual inference algorithm. If we count the number of variables in a p dimensional lag T additive non-linear time series model as $p(T+1)$, then roughly speaking, for longer time series, (80 or more observations per variable), the additive non-linear algorithm outperforms the linear procedure in all situations, including cases where the true model is more complex than the additive non-linear model and cases where the true model is a linear model.

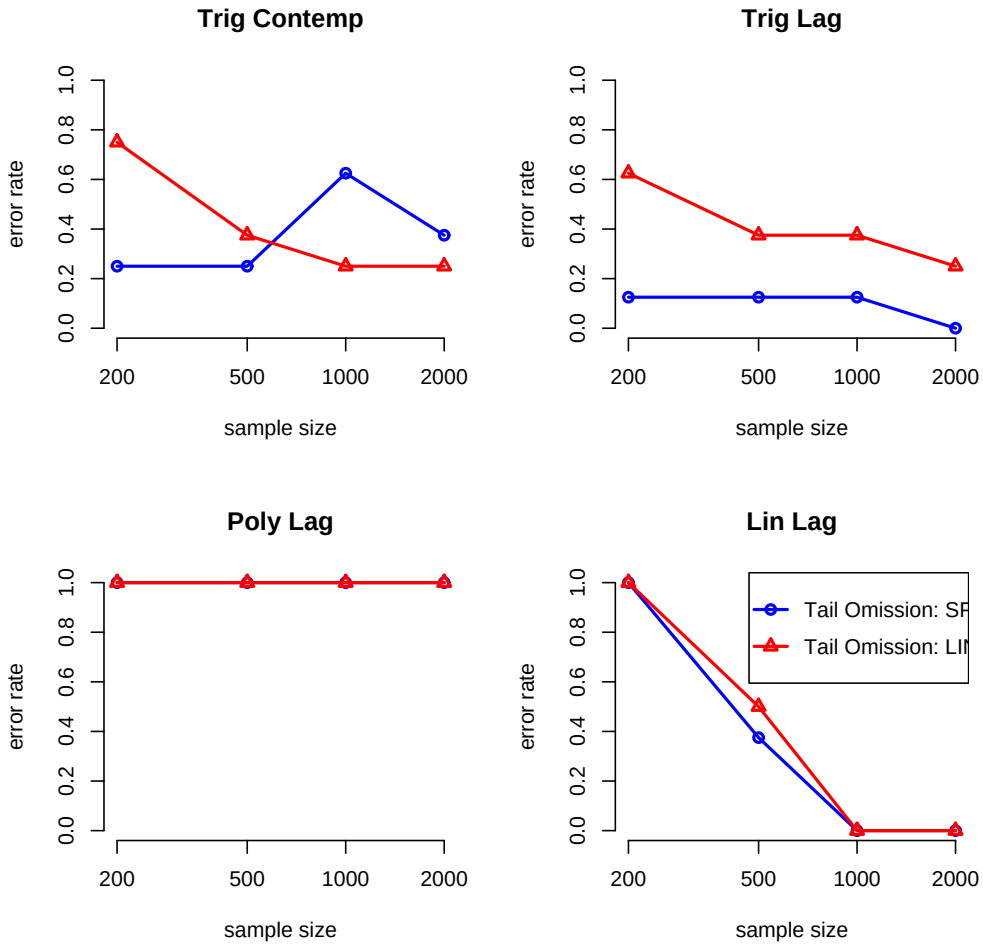

Figure 5: Error rates for edge orientation: Tail

For shorter time series, (less than 40 observations per variable), the additive non-linear model is still better in general, but may be not as good as the linear procedure in some cases. Our suggestion is that, for longer time series always choose the additive non-linear algorithm. For shorter time series, if computational cost is critical, the linear procedure is a reasonable choice; otherwise we still recommend the additive non-linear algorithm, or better yet, try both of them.

5. Case Study: Ocean Climate Indices

To illustrate the application of the additive non-linear causal inference algorithm for nonlinear time series, we use it to study the causal relations among some ocean climate indices.

Climate teleconnections are associations of geospatially remote climate phenomena produced by atmospheric and oceanic processes. The most famous, and first established teleconnection, is the association of the El Nino/Southern Oscillation (ENSO) with the failure of monsoons in India. A variety of associations have been documented among sea surface temperatures (SST), atmospheric pressure at sea level (SLP), land surface temperatures (LST) and precipitation over land areas. Since the 1970s data from a sequence of satellites have provided monthly (and now daily) measurements of such variables, at resolutions as small as 1 square kilometer. Measurements in particular spatial regions have been clustered into time indexed indices for the regions, usually by principal components analysis, but also by other methods. Climate research has established that some of these phenomena are exogenous drivers of others, and has sought physical mechanisms for the teleconnections. We consider here whether constraints on such mechanisms can be obtained by data-driven model selection from time series of ocean indices.

Our data set consists of the following 4 ocean climate indices, recorded monthly from 1958 to 1999, each forming a time series of 504 time steps:

SOI Southern Oscillation Index: Sea Level Pressure (SLP) anomalies between Darwin and Tahiti

WP Western Pacific: Low frequency temporal function of the 'zonal dipole' SLP spatial pattern over the North Pacific.

AO Arctic Oscillation: First principal component of SLP poleward of 20° N

NAO North Atlantic Oscillation: Normalized SLP differences between Ponta Delgada, Azores and Stykkisholmur, Iceland

To check stationarity, we conduct the augmented Dickey-Fuller (ADF) test. ADF tests for all 4 time series reject the null hypothesis that the tested series has a unit root against the alternative that the series is stationary, with p values of the tests smaller than 0.01. As a complementary to ADF tests, we also conduct the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. For all 4 time series, KPSS tests with lag truncation parameter set to 12 fail to reject the null hypothesis that the tested series is (trend) stationary against the unit root alternative, with p values of the tests higher than 0.1. We also plot the autocorrelations for the 4 time series to check if the data satisfies the strong mixing condition (Figure 6). The idea is that, if a time series satisfies the strong mixing condition, its autocorrelation should decrease rapidly as the lag increases. From the plot, the auto correlations of SOI do not decrease as quickly as for other indices, but they become insignificant when the lag is above 12 months.

We assume that the 4 indices are generated from a lag 12 additive non-linear model. The choice of 12 is partly based on the fact that the ocean indices are monthly data. Another concern is that with a length of 504, the data would be too sparse for a model with a much longer lag. As in the simulation study, the R package gam (version 0.97) is used in this analysis. We first remove any linear trend from the data, then, following the causal inference procedure presented in Section 3, derive a causal structure represented by a PAG for the 4 ocean climate indices. Figure 7 gives the learned causal structure.

Because of the relative shorter length of the ocean indices data (10 observations per variable for a lag 12 model), it is worth conducting another inference on the 4 ocean indices using the linear procedure. The resulting causal structure is given in Figure 8.

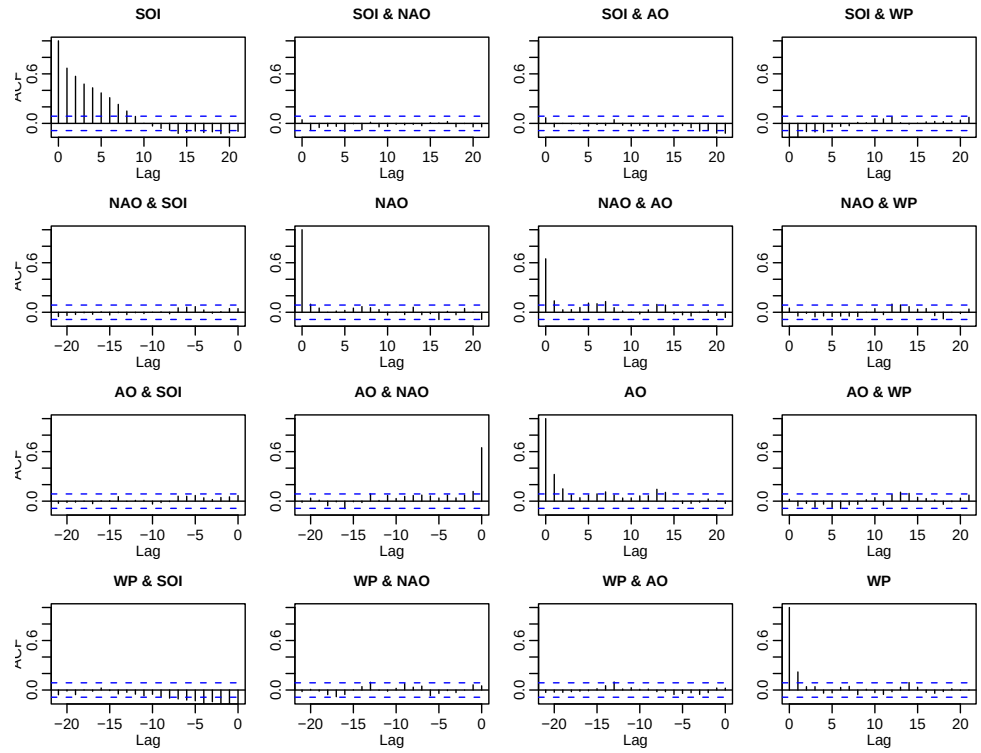

Figure 6: Autocorrelation plot for 4 times series

Without a gold standard, it is hard to say which method gives the more accurate information in this case. But the graph obtained using the linear procedure is likely to miss some nonlinear dependencies. For example, an arrow from SOI_{t-1} to AO_t is present in Figure 7, but absent from Figure 8. It turns out, when regressing AO_t against SOI_{t-1}, AO_{t-1} and NAO_{t-1} using additive model regression, the estimated influence of SOI_{t-1} on AO_t is clearly nonlinear (see Figure 9, where the contribution of SOI_{t-1} to AO_t is plotted as a smooth univariate function of SOI_{t-1}). This is not surprising given the complexity of the processes represented by the ocean climate indices, and illustrates the need of causal inference procedures that can be applied to data generated from nonlinear models.

6. Discussion

Methods of causal inference, first developed in the machine learning literature, have been successfully applied to many diverse fields, including biology, medicine, and sociology (Pearl, 2000; Spirtes et al., 2000). An essential and distinct feature of these methods is that they require comparatively less domain knowledge about the system to be studied.

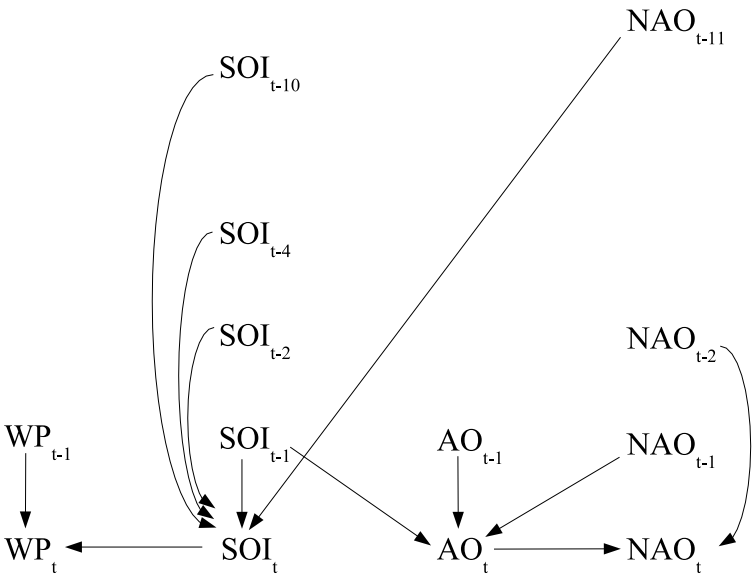

Figure 7: Causal connections among 4 ocean climate indices, using the additive non-linear algorithm

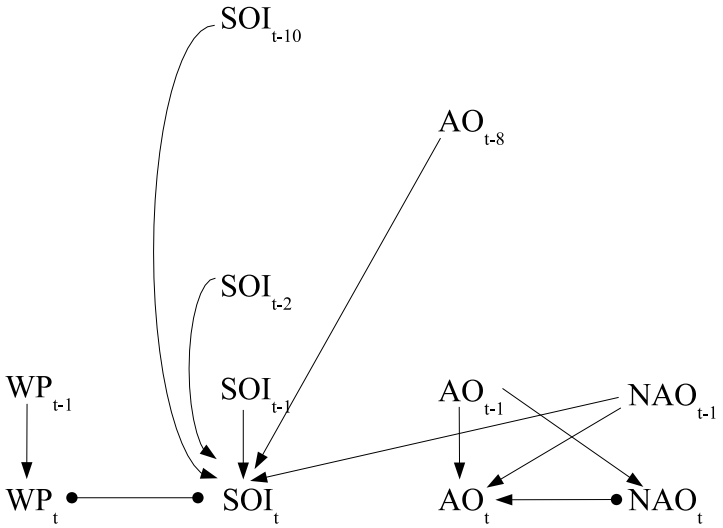

Figure 8: Causal connections among 4 ocean climate indices, using the linear procedure

This study extends the application of causal inference to nonlinear time series data. We present a new procedure that combines semi-automated model search for causal structure with additive

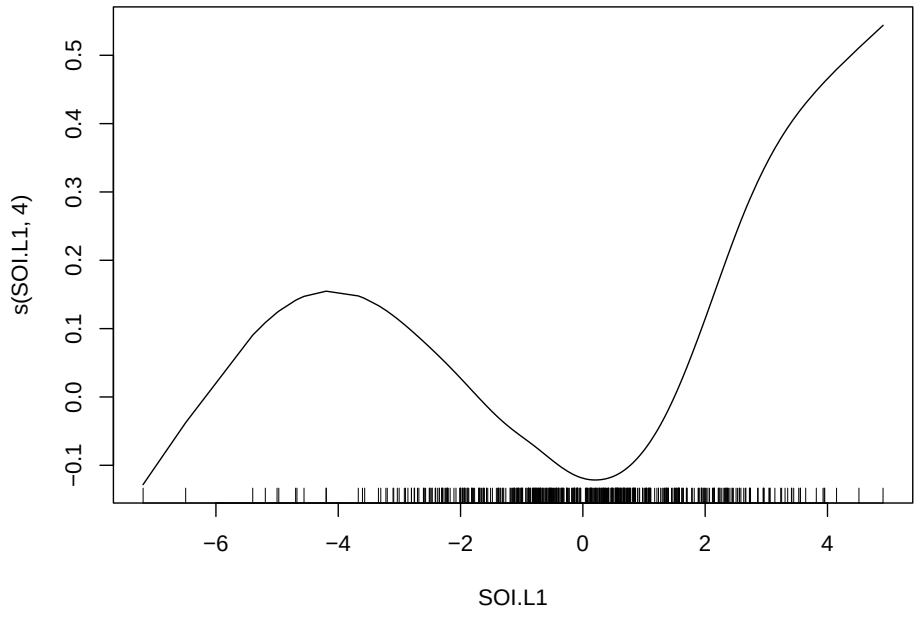

Figure 9: Nonlinear relation between SOI_{t-1} and AO_t

model regression methods. The particular example is to ocean climate indices, but the component procedures have been individually applied to econometric data with some success, suggesting that the criteria for successful application of the joint procedures are statistical and causal rather than domain specific.

Our approach is modular, and its two main components, that is, conditional independence testing and causal model search, could be replaced by other comparable methods. Thus, with appropriate data generated from appropriate mechanisms, related analyses could be conducted under weaker or alternative assumptions. Below we briefly discuss several possible extensions of our method:

6.1 Nonstationary Nonlinear Time Series

In most of this paper we assume that the nonlinear time series are stationary, only because it has been shown that for stationary nonlinear time series data satisfying certain conditions, nonparametric regression is asymptotically consistent. The algorithm and propositions proposed in this paper do not require stationarity. However, to apply our algorithm to nonstationary nonlinear time series data, we must find an efficient regression method to estimate the conditional expectations and conduct conditional independence tests. Cointegration analysis is not suitable for this purpose, because it is mainly designed for and applicable to cointegrated linear time series (Engle and Granger, 1987; Johansen, 1991). However, recent studies on applying nonparametric regression methods to nonstationary time series data (Phillips and Park, 1998; Karlsen and Tjøstheim, 2001; Bandi and Phillips, 2003; Karlsen et al., 2005) seem promising. Not surprisingly, the convergence rate of nonparametric regression for nonstationary time series data may be slower than that of stationary data.

6.2 Feedback Models

The original definition of additive non-linear models in Section 2 does not allow any feedback among contemporaneous variables (see condition C4). To represent mutual influences among contemporary variables, we can remove condition C4 from the original definition. We also have to drop the U terms in Equation 2 because the currently available algorithm capable of handling feedbacks (Richardson and Spirtes, 1999) does not work in the presence of latent common causes. The resulting definition defines a *additive non-linear feedback model*, which, compared to the additive non-linear model, allows feedback, but not latent common causes. Propositions 1, 2, and 3 still hold for the new model, proposition 4 needs some modification:

Proposition 4': Let $\boldsymbol{X}_t^b$ be the set of all contemporary direct causes of X_t^1. Assuming there is no latent common cause, $X_{t-i,j}$ and X_t^1 are dependent conditional on $\boldsymbol{X}_t^b$ and $\boldsymbol{X}^e$ if and only if either $X_{t-i,j}$ is either a direct cause of X_t^1, or a direct cause of a contemporaneous cause of X_t^1.

The only change needed in the causal inference procedure to handle data generated from additive non-linear feedback models is, in step 1(b), that the FCI algorithm should be replaced by a consistent causal inference algorithm capable of outputting cyclic graphs, such as the one proposed in Richardson and Spirtes (1999).

6.3 Score Based Search Procedure

The causal inference procedure presented in Section 3 is constraint based. That is, the procedure requires explicit conditional independence information as input, (although each conditional independence constraint is obtained using a BIC score based model selection procedure). As we mentioned in Section 3, the main advantage of this procedure and its modified version is that they can handle the presence of latent common causes or feedbacks in the contemporaneous causal structure. (Drton et al. 2006 provides a maximum likelihood estimation algorithm that allows the computation of BIC scores for certain types of linear models with correlated error terms, though not for the contemporaneous causal structure of a additive non-linear model.) If we are willing to exclude feedbacks and latent common causes, a simple two-step score based procedure can be used to infer causal information from data generated by additive non-linear models. In the first step, a score based algorithm, such as the GES algorithm (Meek, 1996; Chickering, 2002a,b), is applied to the residuals of additive model regression of contemporaneous variables against all lags to obtain a causal pattern representing a Markov equivalence class π_t of directed acyclic graphs for the contemporaneous variables. In the second step, for each directed acyclic graph G belonging to the Markov equivalence class π_t, we generate a time series causal model M and compute its BIC score in the following way:

- Each contemporaneous variable X_t^i is regressed against its parents in G and all the lagged variables $\boldsymbol{X}^l$. The BIC score method proposed in Huang and Yang (2004) is used to identify the best submodel (with the lowest BIC score s_i). The significant predictors of X_t^i in that best submodel are direct causes of X_t^i in causal model M.

- The BIC score of causal model M is $\sum_i s_i$.

The causal model with the best (lowest) BIC score then is returned as the result of the score based casual inference algorithm.

Acknowledgments

The authors thank the anonymous referees for their helpful comments to improve this paper. This research was completed when the first author was research scientist at Florida Institute for Human and Machine Cognition. The research was supported by NASA contract NC2-1399 to the University of West Florida.

References

H. An and F. Huang. The geometrical ergodicity of nonlinear autoregression models. *Statistica Sinica*, 6:943–956, 1996.

E. Baek and W. Brock. A general test for nonlinear Granger causality: A bivariate model. URL `http://www.ssc.wisc.edu/~wbrock/Baek Brock Granger.pdf`. January 1992.

F. Bandi and P. Phillips. Fully nonparametric estimation of scalar diffusion models. *Econometrica*, 71:241–283, 2003.

D. Bell, J. Kay, and J. Malley. A non-parametric approach to non-linear causality testing. *Economics Letters*, 51:7–18, 1996.

D. Bessler, J. Yang, and M. Wongcharupan. Price dynamics in the international wheat market: Modeling with error correction and directed acyclic graphs. *Journal of Regional Science*, 42: 793–825, 2002.

Z. Cai and E. Masry. Nonparametric estimation of additive nonlinear arx time series: Local linear fitting and projections. *Journal of Econometric Theory*, 16:465–501, 2000.

K. Chan and H. Tong. A note on noisy chaos. *Journal of the Royal Statistical Society B*, 56:301–311, 1994.

R. Chen and R. Tsay. Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, 88:298–308, 1993.

D. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002a.

D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002b.

D. Cline and H. Pu. Geometric ergodicity of nonlinear time series. *Statistica Sinica*, 9:1103–1118, 1999.

Y. Davydov. Mixing conditions for Markov chains. *Theory of Probability and its Applications*, 18: 312–328, 1973.

S. Demiralp and K. Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65:745–767, 2003.

C. Diks and V. Panchenko. A new statistic and practical guidelines for nonparametric Granger causality testing. *Journal of Economic Dynamics and Control*, 30:1647–1669, 2006.

M. Drton, M. Eichler, and T. S. Richardson. Identification and Likelihood Inference for Recursive Linear Models with Correlated Errors. *ArXiv Mathematics e-prints*, August 2006. URL http://arxiv.org/abs/math/0601631v3.

R. Engle and C. Granger. Cointegration and error correction: Representation, estimation, and testing. *Econometrica*, 55:251–276, 1987.

W. Härdle, H. Lütkepohl, and R. Chen. A review of nonparametric time series analysis. *International Statistical Review*, 65:49–72, 1997.

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

C. Hiemstra and J. Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *Journal of Finance*, 49:1639–1664, 1994.

K. Hoover. Automatic inference of the contemporaneous causal order of a system of equations. *Econometric Theory*, 21:69–77, 2005.

P. O. Hoyer, S. Shimizu, and A. J. Kerminen. Estimation of linear, non-Gaussian causal models in the presence of confounding latent variables. pages 155–162, Prague, Czech Republic, 2006.

J. Huang and L. Yang. Identification of non-linear additive autoregressive models. *Journal of the Royal Statistical Society: Series B*, 66:463–477, 2004.

S. Johansen. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, 59:1551–1580, 1991.

H. Karlsen and D. Tjøstheim. Nonparametric estimation in null recurrent time series. *The Annals of Statistics*, 29:372–416, 2001.

H. Karlsen, T. Myklebust, and D. Tjøstheim. Nonparametric estimation in a nonlinear cointegration type model. Working paper, 2005. URL http://www.mi.uib.no/∼karlsen/working_paper/NonlinCoint05.pdf.

C. Meek. *Graphical Models: Selecting Causal and Statistical Models*. PhD thesis, Carnegie Mellon University, Philosophy Department, 1996.

A. Moneta. Graphic models for structural vector autoregressions. Working paper, Laboratory of Economics and Management, Sant'Anna School of Advanced Studies, Pisa, Italy, 2003.

L. Oxley, M. Reale, and G. Tunnicliffe-Wilson. Finding directed acyclic graphs for vector autoregressions. In *Proceedings in Computational Statistics 2004*, pages 1621–1628, 2004.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2000.

P. Phillips and J. Park. Nonstationary density estimation and kernel autoregression. Discussion Paper 1181, Cowles Foundation, Yale University, 1998.

T. Richardson and P. Spirtes. Automated discovery of linear feedback models. In C. Glymour and G. Cooper, editors, *Computation, Causation and Discovery*, chapter 7, pages 253–302. MIT Press, Cambridge, MA, 1999.

P. Robinson. Nonparametric estimation for time series models. *Journal of Time Series Analysis*, 4: 185–208, 1983.

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

P. Spirtes and C. Meek. Learning Bayesian networks with discrete variables from data. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 294–299, 1995.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, New York, NY, second edition, 2000.

C. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13:689–705, 1985.

L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, (forthcoming), 2007.

N.R. Swanson and C.W.J. Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92:357–367, 1997.

D. Tjøstheim and B. Auestad. Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association*, 89:1398–1409, 1994.

Y. Truong and C. Stone. Nonparametric function estimation involving time series. *The Annals of Statistics*, 20:77–97, 1992.

Y. Xia and H. An. Projection pursuit autoregression in time series. *Journal of Time Series Analysis*, 20:693–714, 1999.