Fast and Accurate Causal Inference from Time Series Data

Yuxiao Huang and Samantha Kleinberg
Stevens Institute of Technology
Hoboken, NJ
{yuxiao.huang, samantha.kleinberg}@stevens.edu

Abstract

Causal inference from time series data is a key problem in many fields, and new massive datasets have made it more critical than ever. Accuracy and speed are primary factors in choosing a causal inference method, as they determine which hypotheses can be tested, how much of the search space can be explored, and what decisions can be made based on the results. In this work we present a new causal inference framework that 1) improves the accuracy of inferences in time series data, and 2) enables faster computation of causal significance. Instead of evaluating relationships individually, using only features of the data, this approach exploits the connections between each causal relationship's relative levels of significance. We provide theoretical guarantees of correctness and speed (with an order of magnitude improvement) and empirically demonstrate improved FDR, FNR, and computation speed relative to leading approaches.

Introduction

When testing causal relationships such as "eating chocolate raises blood glucose", we want to know not only that such relationships exist, but when the changes will happen, how large the effect is, and whether there are other factors that interact with the causes. We are often inferring these relationships from massive datasets, such as gene expression microarrays (involving thousands of genes) and high frequency financial market data (such as tick by tick data) and thus need not only methods that can infer these relationships accurately, but ones that can also do so efficiently.

We propose a new approach for causal inference with continuous-valued effects. This improves on prior work by exploiting the connections between each causal relationship's relative levels of significance, and comes with stronger theoretical guarantees while provably lowering the time complexity by an order of magnitude over the state of the art. Finally, we compare the approach to others on simulated data, demonstrating this improved speed and accuracy.

Related work

Current methods have mainly focused on separate aspects of causal inference. A main approach is using Bayesian networks (BNs) (Pearl 2000; Spirtes, Glymour, and Scheines 2000). BNs and their adaptations can handle both discrete and continuous data (Heckerman, Geiger, and Chickering 1995) as well as combinations of the two (Friedman, Geiger, and Goldszmidt 1997), but cannot infer the timing of relationships, and exact structure learning is NP hard (Chickering, Geiger, and Heckerman 1994; Cooper 1990).

Dynamic Bayesian networks (DBNs) (Friedman, Murphy, and Russell 1998; Murphy 2002) extend BNs to time series and similarly aim to find pairwise relationships between variables (rather than more complex combinations). They also only enable inference of relationships with discrete time lags, and remain computationally complex to infer. As a result, heuristics must be used, but these can lead to overfitting and require users to choose many parameters. In contrast, our proposed approach allows efficient exact inference of complex causal relationships with associated windows between cause and effect in time series data.

Granger (1980) causality aims to determine if one time series is predictive of another, after accounting for other available information (often testing this using vector autoregression). The bivariate form uses pairs of variables and is fast, but often finds spurious relationships while the multivariate form is more accurate but extremely computationally complex and in practice not feasible for large datasets. Our approach in contrast is not a regression of one variable on values of others, but rather incorporates the causal significance of one factor when evaluating that of another, and is much less computationally intensive than multivariate Granger.

Recently, Etesami and Kiyavash (2014) used directed information graphs to find linear dynamical graphs, and Peters et al. (2013) used restricted structural equation models to find causal structures from time series where an effect's value is a function of the value of its causes and independent noise. However, the former cannot handle deterministic relationships, and the latter is more complex than our approach ($O(N^2ft)$ versus $O(N^2T)$), where N is the number of variables, f and t the complexity of the user-specified regression method and independence test, and T the length of the time series. Unlike our approach, these methods cannot guarantee the correctness of both relationships and effect size.

Background

An alternative causal inference framework, developed in (Kleinberg 2011), is based on representing causal relationships as logical formulas, enabling inference of complex relationships and their timing (Kleinberg 2012) without prior knowledge. We briefly overview the approach, before discussing several limitations.

A causal relationship here is represented by the following PCTLc (probabilistic computation tree logic with numerical constraints) "leads-to" formula, where c is a discrete variable (e.g. diagnoses) and e a continuous-valued variable (e.g. laboratory tests):

$$c \leadsto^{\geq r, \leq s}_{\geq p} e \neq E[e]. \quad (1)$$

Thus if c is true at timepoint t, then in time window $[t+r, t+s]$, the probability of e deviating from its expected value is at least p. That is, c potentially leads to changes in e's value.

Many factors may fit this definition without being causal, so the major work of this method is separating those that are significant from those that are insignificant or spurious.

Definition 1. Where c is a discrete variable and e a continuous-valued one taking values in $\mathbb{R}$, then c is a *potential cause* of e if: $E[e|c] \neq E[e]$ and c is earlier than e.

Note that there is a time window $[r, s]$ between c and e, as shown in eq. (1), where $1 \leq r \leq s \leq \infty$ and $r \neq \infty$. The above expectations are defined relative to this window.

Definition 2. Where X is the set of potential causes of e, the *causal significance* of c for e, $\varepsilon_{avg}(c, e)$, is:

$$\varepsilon_{avg}(c, e) = \sum_{x \in X \backslash c} \frac{E[e|c \wedge x] - E[e|\neg c \wedge x]}{|X \backslash c|}. \quad (2)$$

This averages the impact of c on the value of e, holding fixed each potential cause of e in turn. Intuitively, an effect of a common cause will make little difference after the true cause is held fixed, unless one common effect consistently precedes the other. Potential causes whose ε_{avg} is greater than a threshold are *ε-significant causes*, with the rest being *ε-insignificant*. For significant causes to be genuine, the data must be stationary and common causes of all pairs of variables must be included. The threshold for ε can be chosen using methods for testing statistical significance while controlling the false discovery rate. The complexity of calculating ε_{avg} with N variables and T timepoints is $O(N^3T)$.

One challenge is when the factors held fixed are correlated and few variables are measured (as this approach assumes there will be many factors to hold fixed). Specifically, eq. (2) uses the difference between $E[e|c \wedge x]$ and $E[e|\neg c \wedge x]$, but there may be causes of e that occur only with $c \wedge x$ or $\neg c \wedge x$, biasing this difference. This occurs in two main cases:

Common cause If y is an unmeasured common cause[1] of c, x and e, and c and x are always earlier than e, then both may seem to cause e. When y is measured, this can still lead to an overestimation of c and x's impact on e if y occurs with both c and x more often than c or x alone. That is, $E[e|c \wedge x]$ would be increased compared to $E[e|\neg c \wedge x]$ and $E[e|\neg x \wedge c]$.

[1] Also known as a latent or hidden variable.

Common effect If c and y are causes of x and e, and x often precedes e, then c, x and y could seem to cause e. If y occurs with x alone more often than with both c and x (as y causes x), then $E[e|\neg c \wedge x]$ may be increased relative to $E[e|c \wedge x]$. This may underestimate the significance of c.

Method

We now propose 1) a new definition of and calculation for causal significance that solves the challenges described, and 2) an approximation that can be used when the exact method cannot. Accuracy and speed of the approach are proven theoretically and later demonstrated empirically. All the proofs are publicly available at: http://www.cs.stevens.edu/∼skleinbe/publications.html.

Our primary contribution is a new measure for causal significance, which aims to isolate the impact of cause c on effect e. Ideally, we want to know what happens to e's value when only c is present and every other cause is absent. However, in practice many causes often co-occur, so the number of such observations is unlikely to be statistically significant.

Instead, we propose that under some assumptions we can treat the significance of a set of causes as a system of linear equations where we then solve for the impact of each cause. This lets us incorporate the significance of each cause and handle cases such as deterministic relationships with few measured variables that previously posed challenges.

Assumptions

We provide stronger guarantees than other approaches, including not only correct inference of a relationship (e.g. c causes e) and its timing (in 1 minute), but also the exact impact (c raises e's value by 6 units). To do this, we rely primarily on the assumptions that:

1. Relationships are linear and additive. That is, the value of a variable at any time is given by the sum of the impact of its causes that are present plus a constant.
2. Causal relationships are deterministic and constant (i.e. c's impact on e is the same every time c occurs). The value of a variable when no cause is present is also constant.
3. All genuine causes are measured.

However, we also demonstrate experimentally that when some or most of the assumptions do not hold, we achieve low false discovery and negative rates.

Causal significance

We now introduce a new definition of causal significance, $\alpha(c, e)$, that measures the average contribution to e's value that comes solely from c. This is done by calculating the average difference between the value of e when only c is present and that when no cause is present.

Definition 3. The *causal significance* of c for e, $\alpha(c, e)$, where X is the set of potential causes of e, and there are relationships of the form $c \leadsto^{\geq r, \leq s} e$ and $x \leadsto^{\geq r', \leq s'} e$ is:

$$\alpha(c, e) = \frac{|T(e|c)|}{N(e|c)} \times (E[e|c \bigwedge_{x \in X \backslash c} \neg x] - E[e| \bigwedge_{x \in X} \neg x]). \quad (3)$$

The difference in expected value gives the difference in e's value due to c alone (when all other causes are absent) and that when all causes (including c) are absent. This accounts for unmeasured and constant background conditions. The difference is multiplied by $|T(e|c)|$, the number of unique timepoints where e is measured in window $[r, s]$ after each instance of c, and divided by the total number of such timepoints $N(e|c)$, to yield the average difference. Potential causes are as in definition 1.

Let $T(v)$ be the set of timepoints where a continuous variable is measured or a discrete one is true. Then:

$$T(e|c) = \bigcup_{c_t} T(e|c_t), \tag{4}$$
$$\text{where} \quad T(e|c_t) = T(e) \cap [t + r, t + s].$$

Similarly, $N(e|c)$ is the total number of timepoints where e could be caused by each instance of c, taking the sum of the size of each set rather than the union. More formally:

$$N(e|c) = \sum_{c_t} |T(e|c_t)|. \tag{5}$$

Note that when multiple instances of a cause c occur such that their windows overlap, then $N(e|c)$ will count the times in the overlap multiple times while $|T(e|c)|$ is the number of unique timepoints. For example, if c's window is $[1, 2]$, c is true at times $\{1, 2\}$ and e is measured at each, then $N(e|c) = 4$ while $|T(e|c)| = 3$. If no windows overlap (say when c is infrequent), $N(e|c)$ reduces to $|T(e|c)|$.

Calculating conditional expectation In practice many causes are correlated, and the negation of all causes aside from c will rarely be observed, leading to a loss of statistical power when calculating this measure from frequencies. Instead, α can be estimated as follows.

While we may not observe the negation of all other causes of e often enough, this difference can be estimated by taking the conditional expectation of e given c and subtracting the component of this caused by other causes of e. This is weighted by the ratio of the number of timepoints in the overlap of c and x's windows, relative to the total number of unique times where c can lead to e. By using this ratio, we can account for cases with many deterministic causes of a single effect. Based on assumptions 1 to 3, the conditional expectation of e given c alone is now:

$$E[e|c \bigwedge_{x \in X \setminus c} \neg x] = E[e|c] - \sum_{x \in X \setminus c} \frac{N(e|c, x)}{|T(e|c)|} \times \alpha(x, e), \tag{6}$$

where:

$$E[e|c] = \frac{\sum_{t \in T(e|c)} e_t}{|T(e|c)|}, \tag{7}$$
$$N(e|c, x) = \sum_{t \in T(x)} |T(e|x_t) \cap T(e|c)|.$$

Similarly, we estimate the expected value of e not due to any potential cause as:

$$E[e| \bigwedge_{x \in X} \neg x] = E[e] - \sum_{x \in X} \frac{N(e|x)}{|T(e)|} \times \alpha(x, e). \tag{8}$$

Under assumptions 1 to 3, eqs. (6) and (8) yield these expectations exactly. Proof is given in sec. A (under corollary A.1) of supplementary material.

Algorithm for efficiently calculating $\alpha(c, e)$

By replacing the expectations with the right-hand side of eqs. (6) and (8), eq. (3) can be written as:

$$\alpha(c, e) = f(e|c) \times (E[e|c] - E[e]) - \sum_{x \in X \setminus c} f(e|c, x) \times \alpha(x, e), \tag{9}$$

where:

$$f(e|c) = \frac{|T(e)| \times |T(e|c)|}{N(e|c) \times (|T(e)| - |T(e|c)|)},$$
$$f(e|c, x) = \frac{N(e|c, x) \times |T(e)| - N(e|x) \times |T(e|c)|}{N(e|c) \times (|T(e)| - |T(e|c)|)}. \tag{10}$$

Note that keeping fixed the timepoints where c and x are true, $f(e|c)$ and $f(e|c, x)$ are then a function of the timepoints where e is measured. When there are no missing data, both are then the same for each e and can be calculated once.

Equation (9) for all $c \in X$ yields the following system of n linear equations and n unknowns:

$$A \times Y = B, \tag{11}$$

where:

$$A = \begin{bmatrix} f(e|c_1, c_1) & \dots & f(e|c_1, c_n) \\ \vdots & & \vdots \\ f(e|c_n, c_1) & \dots & f(e|c_n, c_n) \end{bmatrix}, Y = \begin{bmatrix} \alpha(c_1, e) \\ \vdots \\ \alpha(c_n, e) \end{bmatrix},$$
$$B = \begin{bmatrix} f(e|c_1) \times (E[e|c_1] - E[e]) \\ \vdots \\ f(e|c_n) \times (E[e|c_n] - E[e]) \end{bmatrix}. \tag{12}$$

Here A is an $X \times X$ coefficient matrix for e such that for each $c, x \in X$ the corresponding element is:

$$A_{cx} = f(e|c, x), \tag{13}$$

where $f(e|c, x)$ is defined in eq. (10). We can then calculate $\alpha(c, e)$ (for each $c \in X$) by solving the system. Such a system has a unique solution if A is full rank.

The overall procedure for calculating causal significance for a set of effects is given in algorithm 1.

Correctness

The key contribution of the paper, which claims the correctness of both relationships and exact effect size obtained by algorithm 1, is summarized as follows. Proof is given in sec. A (under theorem A.1) of supplementary material.

Theorem 1. *Under assumptions 1 to 3, if A is full rank, then $\alpha(c, e)$ is exactly the impact of c on e.*

The theorem indicates that, when c is not a genuine cause of e, $\alpha(c, e)$ will be exactly zero.

Algorithm 1 Calculate causal significance

Input:
Continuous-valued time series
Set of effects, $E = \{e_1, \ldots, e_m\}$
Set of potential causes for each $e_i, i \in [1, m]$
$X = \{X_{e_1}, \ldots, X_{e_m}\}$, where X_{e_i} is a set of potential causes $\{c_1, \ldots, c_n\}$ for e_i

Output:
$\alpha(c, e)$ for all $c \in X$ and $e \in E$

1: **for** each e in E **do**
2: Build a system of linear equations based on eq. (11)
3: Calculate $\alpha(c, e)$ for all $c \in X_e$ by solving the system
4: **return** $\alpha(c, e)$ for all relationships

Time complexity

We assume N variables and T timepoints, where $T > N^2$ and each variable is measured (or a value is imputed) for each timepoint. Then, while calculating $\varepsilon_{avg}(c, e)$ is $O(N^3T)$, $\alpha(c, e)$ is $O(N^2T)$, leading to an order of magnitude improvement in speed.

The complexity of building each equation in a system is $O(T)$, with each system having N equations. For N effects, building all equations is $O(N^2T)$. Solving all N systems with direct methods is $O(N^4)$, leading to a total complexity of $O(N^2T + N^4)$. However, as we assume $T > N^2$, the complexity is then $O(N^2T)$. See sec. B (claim B.1) of supplementary material for proof.

When $N < T < N^2$, an alternative method can be used to calculate $\alpha(c, e)$ by making X_e the set of all variables, allowing the use of one equation system rather than N. Under the same assumptions, $\alpha(c, e)$ is still exactly the impact of c, and the complexity of the method remains $O(N^2T)$. Note that although the two methods have the same theoretical complexity, in practice even when $N < T < N^2$, algorithm 1 may be faster as $|X_e|$ can be much smaller than the number of variables.

Approximation

Our guarantees of correctness and speed rely on assumptions that may not hold in all cases. However, we provide an approximate solution to address one main case.

When coefficient matrix A is not full rank, the system of linear equations does not have a unique solution. One possible solution is to find a linearly independent subset of X, X_{lis}, such that its coefficient matrix is full rank, so that the subsystem is guaranteed to have a unique solution.

Now, we propose a greedy method to search for X_{lis} (algorithm 2). The key steps of the method are as follows.

Step 3. Select c_{max} from X to maximize $|E[e|c_{max}] - E[e]|$. By doing this, we attempt to include the most seemingly genuine cause in X_{lis} as early as possible.

Steps 6 to 7. If the coefficient matrix of $X_{lis} \cup c_{max}$ is full rank, then add c_{max} to X_{lis}.

Proof that the coefficient matrix of X_{lis} obtained by algorithm 2 is full rank is given in sec. A (under claim A.1) of supplementary material. Thus a subsystem of eq. (9) for all $\alpha(c, e)$ ($c \in X_{lis}$) is guaranteed to have a unique solution.

Algorithm 2 Search for X_{lis}

Input:
Set of potential causes X
Effect e

Output:
A linearly independent subset of X, X_{lis}

1: $X_{lis} = \varnothing$
2: **repeat**
3: $c_{max} = \arg\max_{c \in X} |E[e|c] - E[e]|$
4: $X = X \backslash c_{max}$
5: A_{lis} = coefficient matrix of $X_{lis} \cup c_{max}$
6: **if** A_{lis} is full rank **then**
7: $X_{lis} = X_{lis} \cup c_{max}$
8: **until** $X = \varnothing$
9: **return** X_{lis}

If X_{lis} includes all genuine causes, then for each $c \in X_{lis}$, $\alpha(c, e)$ is still exactly the impact of c on e. Proof for this claim is the same as for the claim that in general $\alpha(c, e)$ is exactly the impact of c on e and is also shown in the simulated common cause and effect experiment.

The complexity of finding X_{lis} is $O(N^3)$. When $T > N^2$, the complexity of algorithm 1 including this is still $O(N^2T)$. Proof for the two claims are given in sec. B (under claim B.2 and corollary B.1, respectively) of supplementary material.

Experimental results

We compared the proposed method (α for short) with three commonly used methods for causal inference in time series data: that of (Kleinberg 2011) (ε_{avg} for short), dynamic Bayesian networks (DBNs) (Friedman, Murphy, and Russell 1998; Murphy 2002) and bivariate Granger causality (Granger 1980) (Granger for short). For each method, the false discovery rate (FDR, false discoveries as a fraction of all discoveries), false negative rate (FNR, false negatives as a fraction of all negatives) and the run time are reported. Note that each time lag for each relationship is treated as a separate discovery or non-discovery, so if the true window is $[1, 3]$ and an algorithm finds only lag 1, that equates to two false negatives. Run time is the total time for running each method on all datasets sequentially, and does not account for speedups due to parallel computations.

The approaches were tested on multiple simulated datasets (allowing us to evaluate results against ground truth) that incorporate increasingly difficult cases, ranging from simple datasets where all assumptions hold to complex ones where many do not. All the data are publicly available at: http://www.cs.stevens.edu/∼skleinbe/data.html.

Simulated common cause and effect datasets

Methods We generated two datasets with common cause (where variable 1 causing 2 to 4) and common effect (where variables 1 and 2 causing 3 and 4) structures discussed earlier. The data consist of 20 variables (16 of them are noise) and 1000 timepoints. The value of a variable e at timepoint

t, $e(t)$, is given by:

$$e(t) = \sum_{c \in X} \sum_{i=1}^{n} I(c, e), \quad \text{where} \quad n = |T(c) \cap [t-s, t-r]|. \tag{14}$$

Here $T(c)$ is the set of timepoints where c is true, $[r, s]$ the time window of c for e, and $I(c, e)$ the constant impact of c on e. Thus, e's value at t is the sum of the impact of its causes that occurred previously. The time window of all relationships was set as $[r, s] = [1, 3]$, and the impact of each cause was set as $I(c, e) = \pm 5$. The value of variables with no causes were randomly set to zero or one at each time.

DBNs were evaluated with Banjo (Hartemink 2008), with the major parameter settings being: searcherChoice = SimAnneal, minMarkovLag = 1, maxMarkovLag = 3, and maxTime = 1 hour. Banjo requires discrete data, so we used three bins (positive, negative, and zero).

Granger was evaluated with the granger.test function in MSBVAR (Brandt 2012), with time lags in $[1, 3]$. Significant relationships were determined using a p-value cutoff of 0.01.

The overall method and parameter settings of α is the same as for ε_{avg} except for the calculation of causal significance (eq. (11) versus eq. (2)). To calculate the rank of a coefficient matrix and solve the system of linear equations, Gaussian elimination was used. The time window for both methods was set as $[1, 3]$. To determine significant relationships, $z-$values (based on causal significance) were used with $p-$value cutoff of 0.01.

Results Results across both datasets are shown in table 1. The FDR and FNR of α are 0, and the FDR of α is significantly lower than that of ε_{avg} and Granger. Note that DBNs only found relationships at lag 1, instead of 2 and 3. Further, α has the fastest runtime, followed by Granger and ε_{avg}. The runtime of DBNs is a parameter set by the user, as it determines how much of the search space can be explored.

The FDR of α and ε_{avg} differ due to the deterministic relationships with few correlated variables. Take the common cause structure for instance. When using ε_{avg}, each effect was found as a cause of itself and the others. However, when using α, such spurious relationships were not found, and instead their obtained causal significance was 0.

In this experiment, the performance of the method for searching for X_{lis} was also tested. For instance, all 20 variables were originally found as potential causes of variable 2. As the coefficient matrix A was not full rank, X_{lis} had 18 variables including the genuine cause, variable 1. The causal significance of the 17 spurious causes was 0, and that of variable 1 was 5 (its actual impact).

Simulated random datasets

Methods We generated 20 datasets each with a different random causal structure. The data include 20 variables and 1000 timepoints. The differences from the previous experiment are: 1) causality here is probabilistic, with the probability of each cause yielding its effects as 0.5; 2) $r = s$ and was randomly chosen in $[1, 3]$. Thus the data is generated by:

$$e(t) = \sum_{c \in X} \sum_{i=1}^{n} I(c, e) \times \delta, \quad \text{where} \quad n = |T(c) \cap [t-s, t-r]|. \tag{15}$$

Here $P(\delta = 1) = P(\delta = 0) = 0.5$. The mean number (across the 20 datasets) of relationships in each dataset is 41.8, with standard deviation of 4.905.

For DBNs and Granger, their implementation (including the method and threshold for determining significant relationships) are the same as those in the previous experiment. For ε_{avg} and α, thresholds were chosen according to the $z-$values of causal significance, while controlling the local false discovery rate at 0.01 (using the locfdr package (Efron, Turnbull, and Narasimhan 2011)). Time lags in $[1, 3]$ were used for all methods.

Results Results across all 20 datasets are shown in table 1. As in the first experiment, the FDR of α is lower than that of ε_{avg} and much lower than that of Granger, and the FNR is lower than that of DBNs. The run time of Granger is the least, followed by α and ε_{avg}.

Similar to the first experiment, the FDR of ε_{avg} and α differ primarily due to the common cause problem of the former, while DBNs result in the highest FNR. For example, in one dataset ε_{avg} found three spurious relationships between effects of a common cause while α found no spurious relationships in this dataset. Across the 20 datasets, α in fact found only one spurious relationship, and it was one where the relationship is genuine but the time window is incorrect.

Simulated financial datasets

Methods Kleinberg (2012) developed a set of simulated market data, consisting of returns for 25 portfolios with varying causal structures and two 4,000 day observation periods. There are 10 underlying causal structures (sets of causal connections between portfolios) and two different observation periods, yielding a total of 20 datasets. The market model is given by a factor model (Fama and French 1992), so that the return of a portfolio i at day t is:

$$r_{i,t} = \sum_{j} \beta_{i,j} f_{j,t'} + \epsilon_{i,t}, \tag{16}$$

where $\epsilon_{i,t}$ is the sum of the portfolio specific, randomly generated, idiosyncratic terms and all $\epsilon_{k,t-l}$ (where portfolio k is a cause of portfolio i's returns at a lag of l days). All relationships have a randomly generated lag of 1–3 days. The mean number of relationships in each dataset is 20.6, with standard deviation of 12.107.

DBNs were tested as before, using the same parameters and lags. For Granger, many p-values were near zero, so thresholds were chosen based on the F-statistics, while controlling the local false discovery rate at 0.01 (using the locfdr package (Efron, Turnbull, and Narasimhan 2011)).

For ε_{avg} and α, we used the same lags to infer positive causal relationships. That is, the set X only includes c where $E[e|c] > E[e]$. The method and threshold for determining significant relationships are the same as used with the randomly generated datasets.

Table 1: Results on simulated data. Run time is in seconds. *Run time for DBNs is a user-specified parameter.

	Common cause & effect			Random			Finance		
Method	FDR	FNR	run time	FDR	FNR	run time	FDR	FNR	run time
DBNs	0.000	0.006	7200*	0.004	0.025	72000*	0.152	0.013	72000*
Granger	0.488	0.000	23	0.798	0.007	230	0.718	0.015	905
$\varepsilon_{avg}(c, e)$	0.650	0.000	1567	0.053	0.015	68186	0.078	0.012	85678
$\alpha(c, e)$	0.000	0.000	16	0.001	0.006	5088	0.036	0.006	8456

Results Results are shown in table 1. The FDR and FNR of α are lower than those of the other methods.[2] As in the previous experiments, the FDR of ε_{avg} and α differ mainly because of the common cause problem of the former. For instance, in one dataset of 40 relationships with lags ranging from 1 to 3, ε_{avg} found five spurious relationships between effects of a common cause, whereas α found none. The run time of Granger is the least, followed by α and ε_{avg}.

Note that in the financial datasets, the following assumptions do not hold: 1) causal relationships are deterministic, 2) causal impact is constant, and 3) effect's value is constant when no cause is present. However, even though the FDR of α is not zero, we demonstrate that the accuracy is still significantly improved compared to other methods.

Conclusion

Inferring complex temporal causal relationships is important for applications in many areas such as biology and finance, but prior methods either allowed only pairwise relationships, failed to include timing information, or were computationally complex. Other methods have been proposed to address these challenges, but remained computationally complex and faced difficulties with deterministic relationships and datasets with few variables. In this paper, we propose an approach for efficient causal inference from time series data. Unlike previous approaches that use only features of the data and infer causal relationships individually, this method exploits the connections between each relationship's causal significance. The method reduced the time complexity relative to the prior state of the art from $O(N^3T)$ to $O(N^2T)$, while increasing accuracy. In experimental results, our proposed method had both the lowest FDR and FNR on all datasets, while prior approaches faced a tradeoff. Further, the computational speed was reduced significantly compared to the prior state of the art, going from about 26 minutes to 16 seconds in the first dataset, and from about 24 hours to about 2.4 hours in the last. In the future we aim to extend this approach to improve exploration of the hypothesis space (generation of potential causes to be tested), and to identify potential latent variables.

Acknowledgments

This work was supported in part by the NLM of the NIH under Award Number R01LM011826.

[2]Note that the result of ε_{avg} is slightly worse than that reported by (Kleinberg 2011). We attribute this to not using their technique for identifying time windows, and calculating FDR/FNR more strictly for each lag.

References

Brandt, P. 2012. MSBVAR R package version 0.7-2.

Chickering, D. M.; Geiger, D.; and Heckerman, D. 1994. Learning Bayesian Networks is NP-Hard. Technical Report MSR-TR-94-17.

Cooper, G. F. 1990. The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks. *Artificial Intelligence* 42(2–3):393–405.

Efron, B.; Turnbull, B.; and Narasimhan, B. 2011. locfdr: Computes local false discovery rates. R package.

Etesami, J., and Kiyavash, N. 2014. Directed Information Graphs: A Generalization of Linear Dynamical Graphs. In *American Control Conference*.

Fama, E. F., and French, K. R. 1992. The Cross-Section of Expected Stock Returns. *Journal of Finance* 47(2):427–465.

Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian Network Classifiers. *Machine Learning* 29(2-3):131–163.

Friedman, N.; Murphy, K.; and Russell, S. 1998. Learning the Structure of Dynamic Probabilistic Networks. In *UAI*, 139–147.

Granger, C. W. J. 1980. Testing for Causality: A Personal Viewpoint. *Journal of Economic Dynamics and Control* 2:329–352.

Hartemink, A. J. 2008. Banjo: Bayesian Network Inference with Java Objects. http://www.cs.duke.edu/∼amink/software/banjo/.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine learning* 20(3):197–243.

Kleinberg, S. 2011. A Logic for Causal Inference in Time Series with Discrete and Continuous Variables. In *IJCAI*, 943–950.

Kleinberg, S. 2012. *Causality, Probability, and Time*. Cambridge University Press.

Murphy, K. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Dissertation.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Peters, J.; Janzing, D.; and Schölkopf, B. 2013. Causal Inference on Time Series Using Restricted Structural Equation Models. In *NIPS*.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. MIT Press.