A Logic for Causal Inference in Time Series with Discrete and Continuous Variables

Samantha Kleinberg
Columbia University
New York, NY
samantha@dbmi.columbia.edu

Abstract

Many applications of causal inference, such as finding the relationship between stock prices and news reports, involve both discrete and continuous variables observed over time. Inference with these complex sets of temporal data, though, has remained difficult and required a number of simplifications. We show that recent approaches for inferring temporal relationships (represented as logical formulas) can be adapted for inference with continuous valued effects. Building on advances in logic, PCTLc (an extension of PCTL with numerical constraints) is introduced here to allow representation and inference of relationships with a mixture of discrete and continuous components. Then, finding significant relationships in the continuous case can be done using the conditional expectation of an effect, rather than its conditional probability. We evaluate this approach on both synthetically generated and actual financial market data, demonstrating that it can allow us to answer different questions than the discrete approach can.

1 Introduction

Relationships such as "smoking causes lung cancer" or "gene A regulates gene B" can help us decide to quit smoking, or investigate a pathway during drug development, but in many cases it is useful to understand the relationship between the magnitude of the cause and the probability of the effect or explore causes that produce the greatest level of the effect. We may want to know not only whether positive news about a company causes its price to increase, but how much of an increase we can expect, when it will happen, and what other factors are needed. To do this we need to reason about complex relationships that have qualitative, quantitative, and temporal components. Thus far, approaches to causal inference have focused on separate aspects of this problem, with none addressing all areas. Methods based on Bayesian networks [Pearl, 2000] yield compact representations of sparse systems, but neither they nor their temporal extensions, dynamic Bayesian networks [Murphy, 2002], allow for automated representation of and reasoning about relationships more complex than one variable causing another, or involving windows of time. Similarly, Granger causality [1969] allows for only discrete lags between cause and effect and assumes that the relationships are between individual variables. Finally, prior work on inferring complex causal relationships represented as logical formulas identified factors that substantially impact the probability of the effect, but required that all variables be discretized [Kleinberg and Mishra, 2009].

In this work we address the problem of inference when we are interested primarily in the level of the effect. We will show that instead of a difference in probability, a cause's significance for an effect can be assessed using an average difference in conditional expectation. By extending the underlying logic used to represent the relationships, this approach allows for structured representation and automated inference of complex temporal relationships with discrete and continuous components. The result is a succinct representation of the most significant relationships in a set of data. We evaluate this approach empirically on both synthetic and actual financial market data to validate it and assess its practical use.

2 Background

We begin by reviewing the causal inference framework being extended. Earlier work by Kleinberg et al. [2009] introduced a method for causal inference based on probabilistic notions of causality, where the relationship between cause and effect is described in a structured way using probabilistic computation tree logic (PCTL) formulas. To assess whether the formulas satisfied by the data are significant, the average difference each cause makes to the probability of its effect is computed and techniques for multiple hypothesis testing are applied to determine the level at which a relationship should be considered statistically significant [Efron, 2004].

2.1 Probabilistic computation tree logic

We briefly discuss PCTL and refer the reader to [Hansson and Jonsson, 1994] for a more in depth description. Formulas in the logic are defined relative to a probabilistic Kripke structure (also called a discrete time Markov chain (DTMC)), consisting of a set of states, S, a start state s_i, a total transition function $\mathcal{T}$ that gives the probability of transition between all pairs of states and a labeling function L that indicates the propositions from the set of atomic propositions A that are true at each state. In the case of causal inference we do not

normally have or infer these structures but instead test the relationships directly in the data.

There are two types of formulas in PCTL: state formulas that describe properties of individual states, and path formulas that describe properties along sequences of states. These can be defined inductively as follows:[1]

1. Each atomic proposition is a state formula.
2. If f and g are state formulas, so are $\neg f$ and $f \wedge g$.
3. If f and g are state formulas, $0 \leq r \leq s \leq \infty$ and $r \neq \infty$, $fU^{\geq r,\leq s}g$ and $fW^{\geq r,\leq s}g$ are path formulas.
4. If f is a path formula and $0 \leq p \leq 1$, $[f]_{\geq p}$ and $[f]_{>p}$ are state formulas.

The operators in item (2) have their usual meanings. Item (3) describes until (U) and weak-until (W) formulas. First, $fU^{\geq r,\leq s}g$, means that f must be true at every state along the path until g becomes true, which must happen in between r and s time units. The weak-until formula is similar, but does not guarantee that g will ever hold. In that case, f must hold for at least s time units. Finally, in (4) probabilities are added to path formulas to make state formulas. Then the sum of the probabilities of the paths from the state where the path formula holds is at least p. One shorthand that will be useful for representing causal relationships is "leads-to". The formula:

$$f \leadsto^{\geq r,\leq s}_{\geq p} g \equiv AG[f \rightarrow F^{\geq r,\leq s}_{\geq p} g] \tag{1}$$

means that for all paths, from all states, if f is true, then g will become true in between r and s time units with at least probability p. This operator is defined differently relative to traces (where the problem is closer to runtime verification than model checking) as described in [Kleinberg, 2010] and the semantics here.

2.2 Causes as logical formulas

To use this logic for causal inference, Kleinberg et al. [2009] assumed that the system has some underlying structure that is not observed, where the temporal observations (such as stock price movements or longitudinal electronic health records) can be thought of as observations of the sequence of states the system has occupied. In model checking, these sequences are referred to as traces. The standard probabilistic notion of causality, that a cause is earlier than and raises the probability of its effect, can then be translated into PCTL to define potential (*prima facie*) causes.

Definition 2.1. Where both c and e are PCTL formulas, c is a potential cause of e if, relative to a finite trace (or set of traces) or model, the following conditions all hold: the probability of c eventually occurring at some time is greater than zero, the probability of e is less than p and:

$$c \leadsto^{\geq 1,\leq \infty}_{\geq p} e \tag{2}$$

These features are insufficient for identifying causes, since many things (such as common effects of a cause) can occur before and seem to raise the probability of an effect. In order to weed out these spurious causes, one can calculate the average significance of each cause for its effect. The basic premise of this approach is that when testing for spuriousness, one is trying to find whether there are better explanations for the effect. With X being the set of all potential causes of e, the significance of a particular cause c for an effect e is:

$$\varepsilon_{avg}(c, e) = \sum_{x \in X \setminus c} \frac{P(e|c \wedge x) - P(e|\neg c \wedge x)}{|X \setminus c|} \tag{3}$$

Note that the relationships between c and e, and x and e, have time windows associated with them (as in equation (2)), so the conjunctions in (3) refer to instances where e occurs such that either c or x could have caused it (e.g. thinking of each time window as a constraint, both constraints on when e could have occurred are satisfied). This average significance score can be used to partition the potential causes.

Definition 2.2. A potential cause c of an effect e is an ε-insignificant cause of e if $|\varepsilon_{avg}(c, e)| < \varepsilon$.

Definition 2.3. A potential cause c of an effect e that is not an ε-insignificant cause of e is an ε-significant or *just-so* cause of e.

To determine an appropriate value of ε, the problem is treated as one of multiple hypothesis testing, aiming to control the false discovery rate. Assuming many hypotheses are being tested and the proportion of true positives is small relative to this set, methods for empirically inferring the null hypothesis from the data can be applied [Efron, 2004] since the values of ε_{avg} for large scale testing mostly follow a normal distribution, with significant (non-null) values deviating from this distribution [Kleinberg, 2010].

3 Inference of relationships with discrete and continuous components

We now introduce an approach for inferring relationships with continuous-valued effects, and explicitly representing the constraints on continuous valued causes as part of their logical formulas. In the previous section we described evaluating the significance of a cause for its effect using the average difference in probability with each other possible cause of the effect held fixed. When an effect is continuous, we instead want to determine the impact of a cause on the level of the effect, and can do this using the average difference in expected value. For instance, we may want to determine the effect of medications on weight, where it may be difficult to discretize this effect of interest, though the potential causes are naturally discrete variables. In many practical cases the data are noisy and prone to error, making discretization useful, though it is difficult to choose the right partitioning. Further we do not want a generic binning of variables, but rather the ranges of a cause that are most significant for an effect. We propose that by making these part of the logical formula representing the relationship between cause and effect, we can allow for work on automatically reevaluating these ranges after inference in an efficient way (as the significant relationships are a small fraction of the full set tested, constraining the search space).[2]

[1] Lower bounds on time windows do not appear in the original paper by Hansson et al. [1994], but were added by Kleinberg [2010].

[2] In Kleinberg & Hripcsak [2011] we show how time windows can be inferred without prior knowledge of the timing or relation-

The overall approach is to generate a set of logical formulas, determine which are potential causes relative to a set of time series data, then assess the significance of these (partitioning the potential causes into significant/insignificant).

3.1 Representation

Taking inspiration from a similar extension of LTL [Donaldson and Gilbert, 2008], PCTLc (probabilistic computation tree logic with numerical constraints) is introduced to express temporal and probabilistic properties involving continuous and discrete variables.

Syntax

PCTLc formulas are defined relative to a finite set of boolean-valued atomic propositions, A, and variables, V, taking values in $\mathbb{R}$. We will use the convention of referring to members of A using letters from the beginning of the English alphabet and referring to members of V using letters from the end of the alphabet. With $a \in A$, $v \in V$, and f being a function that takes a set of variables and maps these to a value in $\mathbb{R}$:

num ::= $\mathbb{R} \mid v \mid f(v_1, \ldots, v_n)$

State formulas:

$\varphi ::= true \mid a \mid v \bowtie \text{num} \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\psi]_{\geq p} \mid [\psi]_{>p}$

Path formulas:

$\psi ::= \varphi_1 U^{\geq r, \leq s} \varphi_2 \mid \varphi_1 \rightsquigarrow^{\geq r, \leq s} \varphi_2$

where $0 \leq r \leq s \leq \infty$ and $r \neq \infty$; $0 \leq p \leq 1$; and $\bowtie \in \{\geq, >, <, \leq\}$. We initially implement functions, f, for a standard set of mathematical operations ($+, -, *$ and $/$), and note that operators expressing $\vee, \rightarrow$ and weak-until (W) can be derived from the set of operators above.

Semantics

We describe the semantics of PCTLc relative to traces, as this is the most relevant case for causal inference. First we must introduce some terminology. Assume there is a trace T where each timepoint is an observation of the system consisting of truth values for propositions and numerical values for the continuous-valued variables. T could also be a set of traces, though for simplicity we refer to trace T throughout. There is a labeling function $L_a(t)$ that maps timepoints to the atomic propositions true at them, and another $L_v(v, t)$ that maps a timepoint and continuous valued variable to the variable's value at that time.[3] A timepoint t in trace T satisfying formula f is written as $t \models_T f$. A sequence of times (a path) is denoted by $\pi = t_0, t_1, \ldots, t_n$, the subset of π beginning at time i by $\pi^i = t_i, t_{i+1}, \ldots, t_n$, and a particular time t_i in π is $\pi[i]$. The probability of an until (U) formula is the number of timepoints beginning paths that satisfy the formula, divided by the number of times satisfying $\varphi_1 \vee \varphi_2$. That is,

$$\frac{|\{t \in T : \pi^t \models_T \varphi_1 U^{\geq r, \leq s} \varphi_2\}|}{|\{t \in T : t \models_T \varphi_1 \vee \varphi_2\}|} \tag{4}$$

Similarly, the probability of a leads-to ($\rightsquigarrow$) formula is the number of times beginning paths that satisfy the formula, divided by the number of times satisfying φ_1:

$$\frac{|\{t \in T : \pi^t \models_T \varphi_1 \rightsquigarrow^{\geq r, \leq s} \varphi_2\}|}{|\{t \in T : t \models_T \varphi_1\}|} \tag{5}$$

The probability of until and leads-to formulas are dealt with separately as the probabilities must be calculated differently in traces.

The satisfaction relation $\models_T$ is defined as follows.

$t \models_T true$	$\forall t \in T$
$t \models_T a$	if $a \in L_a(t)$
$t \models_T v \bowtie \text{num}$	if $L_v(v, t) \bowtie \text{num}$
$t \models_T \neg\varphi$	if not $t \models_T \varphi$
$t \models_T \varphi_1 \wedge \varphi_2$	if $t \models_T \varphi_1$ and $t \models_T \varphi_2$
$\pi \models_T \varphi_1 U^{\geq r, \leq s} \varphi_2$	if there exists a $j \in [r, s]$ s.t. $\pi[j] \models_T \varphi_2$ and $\pi[i] \models_T \varphi_1, \forall i \in [0, j)$
$\pi \models_T \varphi_1 \rightsquigarrow^{\geq r, \leq s} \varphi_2$	if $\pi[0] \models_T \varphi_1$ and there exists a $j \in [r, s]$ such that $\pi[j] \models_T \varphi_2$
$T \models_T [\psi]_{\geq p}$	if probability of ψ in T is $\geq p$
$T \models_T [\psi]_{>p}$	if probability of ψ in T is $> p$

Truth values of formulas are defined recursively, so to verify these in traces, one may begin by checking the most deeply nested subformulas and build up to the outermost formulas. For example, to check

$$[(a \wedge [v \geq 3]) \rightsquigarrow^{\geq 8, \leq 12} b]_{\geq p}$$

in a trace, we begin by finding the states that satisfy $v \geq 3$, and then label those that also satisfy a with that conjunction. Then, we find timepoints beginning paths that satisfy this initial conjunction where b is true in the specified window afterward. Finally, using the set of states labeled with $(a \wedge [v \geq 3])$ and the paths satisfying the leads-to formula, we can calculate the probability of the formula using equation 5. While we do not discuss the full details here, the complexity of checking a formula f in a trace of length T, where $|f|$ is the size of the formula, is $O(T^2 \times |f|)$. While the approach is straightforward, we can improve the computational complexity in practice by noting that with $f \rightsquigarrow^{\geq r, \leq s} g$, f and g will already be checked and one only needs to test whether the times satisfying f result in g, rather than iterating over all times. Algorithms for checking until and leads-to formulas are described in appendix A.

3.2 Inference

Causal relationships in PCTLc

Using PCTLc, we can express constraints such as $x \leq (y+z)$, and can describe properties such as "elevated glucose for 2-3 months leads to hemoglobin A1C above 8%." Each variable that appears in a state formula must do so with (or as part of) a numerical constraint, so the resulting formula is boolean-valued. Thus with $x \in V$, $[x \geq 5.6] \wedge a \wedge b$ is a valid PCTLc state formula, while $x \wedge a \wedge b$ is not. One could allow unconstrained variables using probability density functions and

ships in a set, and discuss how the algorithm can be applied to the problem of discretization.

[3] We can define such functions relative to a model though that is somewhat more complex, as we want to define the values of variables in each state using constraints (rather than having an observed value).

conditional densities, but this approach can be difficult with noisy data and is computationally complex when a formula contains multiple continuous variables. Many cases of interest also have minimum values for causes to be effective (10mg of aspirin cannot relieve a headache, but 325mg may) so putting together all of the values of the cause will give an unduly low estimate of the cause's significance for the effect.

Conditions for causality

To determine which hypotheses are potential causes of continuous valued effects we must update definition 2.1 in two ways: using expectations rather than probabilities and allowing for negative causes. With binary variables, we can easily view a preventative (something that lowers the probability of an effect, e) as raising the probability of $\neg e$, so that a cause always raises the probability of its effect. Instead, with continuous effects we may have causes that lower the expected value of the effect (such as a drug causing a decrease in weight). To test for potential causality we can calculate the expected value of e, $E[e]$, from the data, then label times where e is greater than this value, $e > E[e]$.[4] Then, a positive causal relationship can be represented with the following leads-to formula:

$$c \leadsto^{\geq r, \leq s}_{\geq p} [e > E[e]]. \quad (6)$$

Definition 3.1. Where c is a PCTLc formula and e is a continuous-valued variable taking values in $\mathbb{R}$, c is a potential cause of e if, with c being earlier than e:

$$E[e|c] \neq E[e]. \quad (7)$$

This says that when c occurs we can expect a different value of e than we would otherwise.[5] Note that there is a window of time $[r, s]$ between c and e, where $1 \leq r \leq s \leq \infty$ and $r \neq \infty$. We omit the temporal subscripts, but note that the expectations are defined relative to these. With a set of time series data the expectation is calculated using frequencies ($\#(x)$ denotes the number of timepoints where x holds):

$$E[e|c] = \sum_y y \frac{P(e = y, c)}{P(c)} = \sum_y y \frac{\#(e = y, c)}{\#(c)} \quad (8)$$

Significance of causes

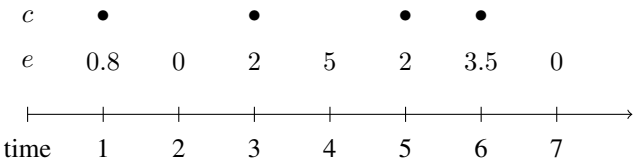
To find which of a set of possible causal relationships are significant, we need to assess how much of a difference the causes make to the value of their effects. We want not just the average effect that results from the cause, but the average effect conditioned on other possible explanations. The significance of a potential cause c (defined as in 3.1) for an effect, e, where X is the set of prima facie causes of e is measured by:

$$\varepsilon_{avg}(c, e) = \sum_{x \in X \setminus c} \frac{E[e|c \wedge x] - E[e|\neg c \wedge x]}{|X \setminus c|}. \quad (9)$$

The definitions of ε-significant and ε-insignificant remain as in 2.2 and 2.3, but with ε_{avg} calculated as in equation (9). This measure retains the important properties of the previously defined ε_{avg}: with a large number of insignificant relationships it will be normally distributed, and factors spuriously associated with an effect will have small values of it. As before, some ε-insignificant relationships may be spurious (e.g. they can be explained by a common cause) or may simply be very small in magnitude, in which case they are of little use for our purposes. Note that there is no difficulty in handling multiple causes of an effect that vary in significance and that this approach allows us to correctly identify a common cause of two effects as such.

Difference from discrete case

Let us look at an example illustrating how the results of this approach can differ from those obtained using the probability of a discretized effect. Below is a short sequence of observations of a continuous variable e, and a discrete event c.



To determine if c is a potential cause of e in exactly 1 time unit, we test whether the expected value of e and the conditional expectation of e given c differ. Here $E[e]$ is 1.9, while $E[e|c]$ is 2.125, so c increases the expected value of e and is a potential cause of it. If instead we discretize e using a common approach, calling values below the expected value "false" and above "true", we find that $P(e)$ is $4/7$ (≈ 0.57), and $P(e|c) = 0.5$, so c does not raise the probability of e and would not be considered as a potential cause of it. While we omit the subscripts, this is the probability of e being true in exactly one time unit after c being true. In this case it may be that c is part of a significant cause, $c \wedge d$ where c and d alone cannot cause e. If we have unmeasured variables or missing data, as we do here since d is not included, it can be difficult to find c's relationship to e. Using the expected value allows us to find factors that make a large but less consistent difference to their effects, while the probability difference requires a consistent impact. We demonstrate this experimentally in Section 4.2 where we find relationships that are extremely significant during only part of a financial time series studied.

4 Experimental results

One area where we are particularly interested in the level of effects is in finance. For risk management and portfolio rebalancing, we want to know not only what causes a price increase, but that one cause leads to an expected gain of 15% with probability 0.2, while another leads to a gain of 1% with a probability of 0.8. To test the approach in this area and understand how it compares to methods that discretize the variables, we first evaluate it on simulated returns where the embedded relationships are known and false discovery rates can be calculated, and then apply it to actual market data.

4.1 Synthetic financial time series

Kleinberg et al. [2010] developed a dataset simulating daily stock market returns with various types of causal relation-

[4] We may also test more specific properties such as $x \leq E[e] \leq y$, labeling the timepoints where this is true and proceeding as above.

[5] One could replace $\neq$ in the definition with $>$ to stipulate only positive prima facie causes and $<$ for negative ones.

ships embedded in the data. Using a factor model based on the Fama-French daily factors [1993], that work created a market consisting of 25 portfolios for two 3001 day time series with five types of causal influence and one scenario with no embedded causality (yielding a total of 12 data sets). The return of a portfolio i at time t is given by:

$$r_{i,t} = \sum_j \beta_{i,j} f_{j,t'} + \epsilon_{i,t} \tag{10}$$

where the return of factor j at time t' is $f_{j,t'}$, $f_{j,t'} \in \mathbb{R}$ and $\epsilon_{i,t}$ is a random variable with mean zero. The null case is $t' = t-3$ for all factors and portfolios. There were two cases with shifted factors: one with half the stocks influenced by factors at $t' = t-1$, and another where half the stocks had their factors lagged individually by a random value in $[0, 3]$. Data was generated for these three scenarios, and then the same scenarios with some direct relationships between portfolios added (captured through dependence of one portfolio on the residual term of another at the previous timepoint). In addition to the direct influences, we should find that stocks responding to earlier factors cause those responding later as the factors are not independent of the portfolios.

We compare our results to those from [Kleinberg *et al.*, 2010] and the application of Granger [1969] causality from that work. The ε_{avg} values remained normally distributed in the absence of embedded causal relationships, ensuring that when there are no causal relationships we do not infer any. In order to determine which ε_{avg} values were statistically significant, we initially applied the same approach as in [Kleinberg, 2010], inferring the empirical null using the locfdr package [Efron *et al.*, 2008] and controlling the local false discovery rate at 0.01. The false discovery rate (FDR, proportion of spurious causes called significant out of all causes called significant) and false negative rate (FNR, proportion of true causes called insignificant out of all causes called insignificant) for our approach over all datasets were: 0.144 and 0.039 while those of [Kleinberg *et al.*, 2010] were: 0.077 and 0.042 and the MSBVAR implementation of Granger causality were: 0.654 and 0.086. In the datasets that lag factors for half the stocks, there are a large number of true positives, violating one of the assumptions of methods for empirical inference of the null hypothesis. Using the theoretical null hypothesis (that the null values are given by $N(0, 1)$) instead leads to a low FDR, 0.030, at the expense of a somewhat larger FNR, 0.100. As shown in figure 1, these cases have bimodal distributions, with a clear separation between the null and non-null values. When we use this separation to choose the threshold (where $f(z)$ is at its minimum between the peaks), the method described here outperforms the approach based on discrete values, with an FDR of 0.003 compared with 0.010 for the discrete method (in this case both have the same FNR, 0.048). Thus using continuous data does better than the discretized data when we do not use the empirical null to choose the threshold.

4.2 Actual market returns

In addition to evaluating the proposed approach on a set of simulated data, it was also applied to actual daily market data.

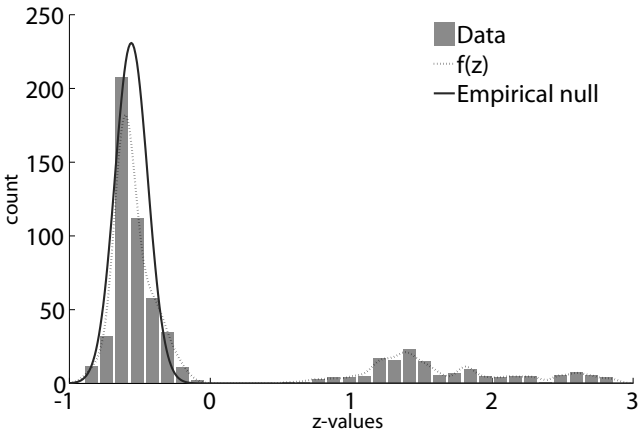


Figure 1: Histogram of z-values (calculated as the number of standard deviations of an ϵ_{avg} from the mean) for one simulated financial market data set where half the stocks have their factors shifted by the same amount.

Wharton Research Data Services (WRDS) was used to obtain daily returns for the same time period and set of stocks as in [Kleinberg *et al.*, 2010], yielding all 252 trading days in 2007 and 286 stocks from the S&P 500. Relationships between pairs of stocks were tested with a time window of exactly 1 day (as it is believed that there is a sharp drop off in influence after that time). Interestingly, a number of the top relationships found involved a price decrease in stocks (such as Schlumberger Ltd. and Smith International Inc.) causing a price increase in others (particularly Goldman Sachs Group Inc.). Unlike the synthetic case, where the true relationships had the highest ε_{avg} values in both methods, and the biggest difference was in estimation of the null hypothesis, results differed considerably on the actual market data.

Here many more significant relationships were identified: 915 versus 27 for the discretized data. We believe that the reason for this is likely the non-stationary nature of the real data. While the simulated market was governed by the same relationships throughout the entire time series, this is not necessarily the case for the actual market, where regulations and partnerships between companies change constantly and there is much hidden information. Discretized data seems to favor inference of relationships that persist across time regardless of magnitude, while the continuous data can allow us to find very large effects that may be true for only part of the time series. This is useful for analysis of data in a changing market, since we want to know what the important relationships will be going forward, not just what they were in the past.

In order to verify this intuition we examined these results in detail and conducted further simulations. First, there were many relationships satisfied for only part of the time series. That is, a price increase was insignificant at the beginning of the time series before increasing dramatically in the later part. In the case of Goldman Sachs Group Inc. (GS) increasing after each time Schlumberger Ltd (SLB) decreased, the expected value of an increase was 0.5 points for the first 150 days, and 2.2 during the last 102 days, increasing further to

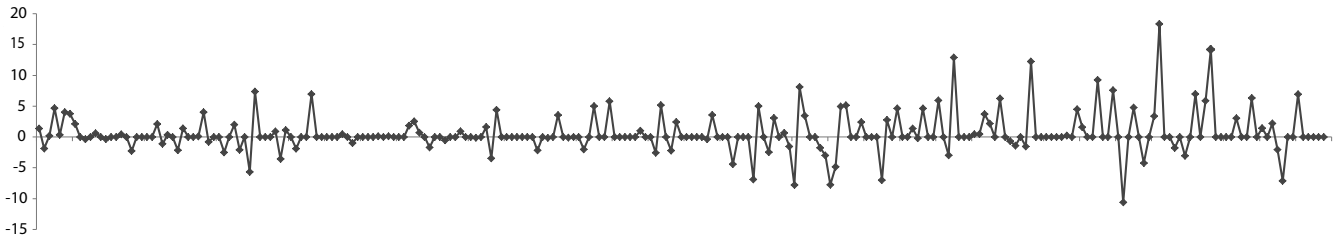


Figure 2: Price changes in GS each day after SLB is down the prior day are shown for one year. The expected value of an increase in GS is much larger during the second half of the time series than in the first half. Values of zero are cases where SLB was up the previous day.

3.0 for the final 52 days. Figure 2 shows the price change in GS after each instance of SLB decreasing. While it may appear visually that this is due to more instances of SLB being down in the last half, this is not supported by the numerical analysis. We further verified this by combining data from the simulated series such that the first two-thirds of the data came from one distribution (with no embedded causality) and the last one-third from another (with a few relationships between portfolios). Here we were again able to correctly infer some relationships that were true for only a short time, while making the same amount of false discoveries as in the stationary case.

5 Related work

5.1 Logic

A number of logics have recently been developed for reasoning about relationships involving qualitative, quantitative, temporal and probabilistic components. One of the most similar to ours, PLTLc, was developed for studying biochemical networks where there is a need to represent the concentration of variables [Donaldson and Gilbert, 2008]. The primary difference between it and our approach is that PLTLc uses standard modal operators such as "next" and "finally" to describe when properties hold, but does not allow quantification of time. In contrast, PCTLc allows us to reason about the amount of time a property holds, or how long it will take for a property to hold after a condition is met. Most seemingly similar logics, such as [Kanazawa, 1991], relate primarily to planning problems, where causality amounts to changing the truth values of variables. Instead of deduction from a set of rules or finding temporal patterns we aim to infer causal relationships from data. The recently developed annotated probabilistic temporal (APT) logic allows for representation of properties such as "after a patient takes drug A they will develop symptom B with a probability between 10 and 20% within 2 months" [Shakarian *et al.*, 2011]. However this approach is closer to frequent pattern mining, making it likely to find relationships between effects of a common cause (e.g. yellowed fingers being followed by lung cancer) and including unnecessary but frequent features. In contrast we would find that smoking causes both yellowed fingers and lung cancer. While a pattern could be useful for planning, it is not sufficient for intervening on a system and does not provide any new knowledge about the system.

5.2 Causal Inference

The most similar approaches to ours in the causal inference literature are extensions to graphical models and methods based on Granger causality. Graphical model approaches infer directed acyclic graphs called Bayesian networks (BNs), where nodes represent variables and edges between them indicate conditional dependencies [Pearl, 2000]. BNs have primarily been applied to discrete variables, but there are extensions to the case of continuous variables and mixtures of continuous and discrete variables. One of the earliest methods, introduced by Lauritzen [1989], showed that for a set of continuous variables that are assumed to have linear conditional Gaussian distributions, the influence of parent nodes on their children can be represented as a shift in their means. A similar approach is taken for continuous-valued descendants of discrete parents, though other approaches are needed for modeling discrete effects of continuous parents [Murphy, 1999]. The resulting graph, called a hybrid Bayesian network, has been used for applications such as studying gene expression data [Friedman *et al.*, 2000]. However, these adaptations face many of the same limitations as BNs since they cannot represent and infer relationships involving complex sequences of factors and exact inference is not possible, requiring approximations [Lerner and Parr, 2001].

Dynamic Bayesian networks extend BNs to reason about time by using one BN for each time slice and connecting BNs across time slices to indicate how variables at one time influence those at another. Hybrid DBNs and the recently introduced continuous dynamic [Grzegorczyk and Husmeier, 2009] and heterogeneous DBNs [Dondelinger *et al.*, 2010] (which allow for non-stationary time series) extend DBNs for inference with continuous-valued variables. However, like DBNs they cannot reason about windows of time or complex relationships in a structured way. Granger causality [Granger, 1969], tests whether lagged values of one time series are informative about another, usually using regression-based methods. These can handle continuous-valued variables and while there are some multivariate extensions [Barrett *et al.*, 2010], we are not aware of any versions that go beyond individual lags (allowing a window of time as in our approach) or that allow for the complex types of relationships that can be represented in our logic.

6 Conclusion

Causal inference with a mixture of discrete and continuous variables observed over time is important for many areas, including finance and the analysis of electronic health records. We have proposed an extension of PCTL called PCTLc for representing probabilistic relationships involving both discrete and continuous valued variables (as well as numerical and temporal constraints), and a new method for evaluating the significance of causes of continuous-valued effects using conditional expectations. While ours are not exhaustive criteria for causality (not all genuine causal relationships will meet the criteria set forth), this framework allows us to ask different types of questions than have previously been possible, in an efficient way while minimizing false discoveries. We evaluated the method on both simulated and actual financial market data, finding that in the simulated case where there are no hidden variables, the time series is stationary, and relationships are not more complex than pairwise, there is minimal difference between discretization and using the continuous values (though the FDR with continuous data was somewhat lower than that with discretized data). However, in actual market data we are able to make many more inferences and find relationships that were undetected by other approaches. In particular, we are able to make discoveries in non-stationary time series where relationships may only be true for part of the period under study.

Acknowledgements

This material is based upon work supported by the NSF under Grant #1019343 to the Computing Research Association for the CIFellows project. This work has also been funded in part with Federal funds from the NLM, NIH, DHHS, under Contract No. HHSN276201000024C.

A Algorithms

To check a probabilistic until formula $[fU^{\geq r,\leq s}g]_{\geq p}$ in a trace T, we can iterate over the states satisfying either f or g, rather than the full trace,

$$\text{until-P}(f,g,r,s) = (F \leftarrow \{t : t \models_T f\} \quad G \leftarrow \{t : t \models_T g\} \quad \frac{|\{t \in F \bigcup G : \text{until}(f,g,r,s,t)\}|}{|F \bigcup G|})$$

where the satisfaction of a path beginning at time t of the until formula is given by:

$$\text{until}(f,g,r,s,t) = \begin{cases} true & \text{if } t \models_T g \wedge r \leq 0 \\ false & \text{if } t \not\models_T f \vee t = |T| \vee s = 0 \\ \text{until}(f,g,r\text{-}1,s\text{-}1,t\text{+}1) & \text{otherwise} \end{cases}$$

Leads-to formulas can be checked similarly using:

$$\text{leadsto-P}(f,g,r,s) = (F \leftarrow \{t : t \models_T f\} \quad \frac{|\{t \in F : \text{leadsto}(f,g,r,s,t)\}|}{|F|})$$

where the path formulas are checked using:

$$\text{leadsto}(f,g,r,s,t) = \begin{cases} true & \text{if } (t \models_T f) \wedge (\pi^{t+r} \models_T true U^{\geq 0,\leq s-r} g) \\ false & \text{otherwise} \end{cases}$$

Correctness of the algorithms follows from the definitions and is easily shown inductively (proofs proceed nearly identically to those in [Kleinberg, 2010]).

References

[Barrett *et al.*, 2010] A.B. Barrett, L. Barnett, and A.K. Seth. Multivariate Granger causality and generalized variance. *Physical Review E*, 81(4):41907, 2010.

[Donaldson and Gilbert, 2008] R. Donaldson and D. Gilbert. A model checking approach to the parameter estimation of biochemical pathways. In *Computational Methods in Systems Biology*, pages 269–287. Springer, 2008.

[Dondelinger *et al.*, 2010] F. Dondelinger, S. Lebre, and D. Husmeier. Heterogeneous continuous dynamic bayesian networks with flexible structure and inter-time segment information sharing. In *ICML*, pages 303–310, 2010.

[Efron *et al.*, 2008] B. Efron, B. Turnbull, and B. Narasimhan. locfdr R package, 2008.

[Efron, 2004] B. Efron. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, 99(465):96–105, 2004.

[Fama and French, 1993] E.F. Fama and K.R. French. Common Risk factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[Friedman *et al.*, 2000] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

[Granger, 1969] C.W.J. Granger. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3):424–438, 1969.

[Grzegorczyk and Husmeier, 2009] M. Grzegorczyk and D. Husmeier. Non-stationary continuous dynamic Bayesian networks. *Advances in Neural Information Processing Systems (NIPS)*, 22:682–690, 2009.

[Hansson and Jonsson, 1994] H. Hansson and B. Jonsson. A Logic for Reasoning about Time and Reliability. *Formal Aspects of Computing*, 6(5):512–535, 1994.

[Kanazawa, 1991] K. Kanazawa. A logic and time nets for probabilistic inference. In *Proceedings, Ninth National Conference on Artificial Intelligence*, pages 360–365, 1991.

[Kleinberg and Hripcsak, 2011] S. Kleinberg and G. Hripcsak. When and Why: adaptive inference of temporal causal relationships. Under review, 2011.

[Kleinberg and Mishra, 2009] S. Kleinberg and B. Mishra. The Temporal Logic of Causal Structures. In *Proc. of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 303–312, 2009.

[Kleinberg *et al.*, 2010] S. Kleinberg, P. Kolm, and B. Mishra. Investigating Causal Relationships in Stock Returns with Temporal Logic Based Methods. *ArXiv e-prints*, June 2010.

[Kleinberg, 2010] S. Kleinberg. *An Algorithmic Enquiry Concerning Causality*. PhD thesis, New York University, 2010.

[Lauritzen and Wermuth, 1989] S.L. Lauritzen and N. Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, 17(1):31–57, 1989.

[Lerner and Parr, 2001] U. Lerner and R. Parr. Inference in hybrid networks: Theoretical limits and practical algorithms. In *Proc. of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 310–31, 2001.

[Murphy, 1999] K Murphy. A variational approximation for Bayesian networks with discrete and continuous latent variables. In *Proc. of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 457–466, 1999.

[Murphy, 2002] K. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkley, 2002.

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

[Shakarian *et al.*, 2011] P. Shakarian, A. Parker, G. Simari, and VS Subrahmanian. Annotated Probabilistic Temporal Logic. *Transactions on Computational Logic*, 12(2), 2011.