On Causal Discovery from Time Series Data using FCI

Doris Entner[1] and Patrik O. Hoyer[1,2]

[1] HIIT & Dept. of Computer Science, University of Helsinki, Finland

[2] CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

We adapt the Fast Causal Inference (FCI) algorithm of Spirtes et al. (2000) to the problem of inferring causal relationships from time series data and evaluate our adaptation and the original FCI algorithm, comparing them to other methods including Granger causality. One advantage of FCI based approaches is the possibility of taking latent confounding variables into account, as opposed to methods based on Granger causality. From simulations we see, however, that while the FCI based approaches are in principle quite powerful for finding causal relationships in time series data, such methods are not very reliable for most practical sample sizes. We further apply the framework to microeconomic data on the dynamics of firm growth. By releasing the full computer code for the method we hope to facilitate the application of the procedure to other domains.

1 Introduction

One of the fundamental goals in empirical science is to discover causal relationships. The most reliable approach towards this goal is performing controlled experiments; regrettably, such experiments are not always possible, for technical or ethical reasons, or high cost. Thus, scientists often seek (preliminary) causal inferences from non-experimental data.

Such data often come in the form of multivariate time series. For instance, economists study the evolution of variables such as GDP, unemployment, and interest rates, while biologists may look at the population dynamics of a set of species. In these cases, a number of variables have been measured repeatedly over time, and the goal is often to understand the myriad ways in which the variables interact.

In time series analysis, most approaches to causal inference are based on Granger causality (Granger, 1969). The basic idea is to, for each variable, identify the set of other variables whose past values are necessary and sufficient for optimal prediction. Unfortunately, if unmeasured variables directly affect two or more of the observed variables, one cannot directly interpret the results of such an analysis.

There exist, however, inference procedures which are asymptotically correct even in the presence of hidden variables. We adapt and evaluate the Fast Causal Inference (FCI) method of Spirtes et al. (2000), originally developed for the analysis of non-temporal variables, to time series data. We demonstrate the statistical behavior of the adapted procedure using numerical simulations, and compare it to other recently developed causal inference methods. In addition, we apply it to a real-world dataset on the dynamics of firm growth. We hope, by releasing the full implementation of the method, to facilitate the further development and adoption of this procedure by researchers in a variety of fields.

2 'FCI' in a nutshell

2.1 Problem definition

A directed graph $\mathcal{G}$ is a pair (V, E) where $V = \{1, \ldots, M\}$ is a finite set of vertices and $E \subseteq V \times V$ is the set of directed edges between the vertices. If $(i, j) \in E$ then we write $i \rightarrow j$ and say that i and j are *adjacent*, j is a *child* of i, i is a *parent* of j, and write $i \in \pi(j)$, the parent set of j. A *directed path* from i to j is a sequence of one or more edges $i \rightarrow \ldots \rightarrow j$ where all edges along the path point towards j. In a *directed acyclic graph* (DAG) there are no directed paths from any vertex to itself. If there exists a directed path from vertex i to j, or if $i = j$, we say that i is an *ancestor* of j, and j is a *descendant* of i.

DAG structures are often used to model data-generating processes, whereby each vertex i of $\mathcal{G}$ represents one random variable X_i. Furthermore,

we associate with each vertex i a conditional distribution $P(X_i \mid X_{\pi(i)})$ representing the mechanism by which X_i is generated conditional on the values of its parents in the graph. Such a model is causal if, when a given variable X_i is intervened on and *set* to a specific value, the post-interventional distribution is represented by the same model but with all edges *into* node i removed (Pearl, 2000).

If the data is generated as described above, with DAG $\mathcal{G}$, the joint distribution $P(\mathbf{X})$ (where $\mathbf{X} = (X_1, \ldots, X_M)$) contains independencies related to the structure of $\mathcal{G}$. This is embodied in the graphical criterion of *d-separation* (Pearl, 1988); if vertices i and j are d-separated given some subset $K \subseteq \mathbf{V} \setminus \{i, j\}$ then in the distribution $P(\mathbf{X})$ we necessarily have that X_i is independent of X_j given the variables $\{X_k : k \in K\}$, which we write $X_i \perp\!\!\!\perp X_j \mid X_K$. This is known as the *Markov* condition. If, furthermore, *all* independencies in $P(\mathbf{X})$ are entailed by d-separation in $\mathcal{G}$, then we say that the distribution $P(X)$ is *faithful* to the graph $\mathcal{G}$.

If the data-generating process results in a distribution faithful to the underlying $\mathcal{G}$, constraint-based search algorithms can be used to infer various aspects of $\mathcal{G}$. In general, causally different graphs can result in the same set of independencies in the data, a phenomenon known as *Markov equivalence*. However, in many cases the set of all graphs which are consistent with the observed pattern of independencies in the data have some features in common, and such features can then be inferred. For *causally sufficient* systems, where all the X_i, $i = 1, \ldots, M$, have been observed, the Markov equivalence classes are particularly simple (Spirtes et al., 2000).

2.2 MAGs, PAGs, and FCI

For the more involved task of finding equivalence classes in causally insufficient systems Spirtes et al. (2000) developed the FCI algorithm. Its output graph contains partial information about *ancestral* relationships among the observed variables and is thus termed a *partial ancestral graph* (PAG).

Towards this end, we need the following definitions (Richardson and Spirtes, 2002). A *mixed graph* is a graph that contains three types of edges: undirected (—), directed (→) and bi-directed (↔). (As we exclude selection bias we will not be dealing with undirected edges so all our subsequent defini-

tions are conditional on this.) There can be at most one such edge between any given pair of vertices, and no edge can connect a vertex to itself. The terms parent, child, directed path, ancestor and descendant are defined as for DAGs. Additionally, if $i \leftrightarrow j$ then i is a *spouse* of j. An *ancestral graph* is a mixed graph for which there is no vertex i which is an ancestor of any of its parents nor any of its spouses.

Ancestral graphs can represent systems derived from DAGs but in which a subset of the variables have been hidden. A graphical separation property termed *m-separation* can be defined, mirroring d-separation for DAGs, such that m-separation in an ancestral graph corresponds to independencies between the observed variables in the distribution. A *maximal ancestral graph* (MAG) is an ancestral graph such that for every pair of variables $\{X_i, X_j\}$ there is an edge between i and j if and only if there does not exist a set $K \subseteq \mathbf{V} \setminus \{i, j\}$ such that $X_i \perp\!\!\!\perp X_j \mid X_K$. See (Richardson and Spirtes, 2002).

If two MAGs entail the same set of conditional independencies they are said to be Markov equivalent. Thus, a PAG describes an equivalence class of MAGs, and is a graph with edges having three kinds of edge marks: arrowtails, arrowheads, and circles, in any combination. A PAG $\mathcal{P}$ of an equivalence class fulfills that $\mathcal{P}$ has the same adjacencies as any member of the equivalence class and every non-circle mark is present in all members of the equivalence class. As we exclude selection bias (and hence undirected edges), in our case a PAG can contain the following types of edges: →, ↔, ∘→, and ∘—∘. An edge $X_i \rightarrow X_j$ is interpreted as X_i being an ancestor of X_j (in the underlying DAG), and X_j not being an ancestor of X_i. If $X_i \leftrightarrow X_j$ then neither variable is an ancestor of the other. The circle mark represents cases where, in different members of the equivalence class, both arrowtails and arrowheads are present at this end; hence, based on independencies alone, it is undecided whether that variable is an ancestor or non-ancestor of the other variable.

Finally, the FCI algorithm (Spirtes et al., 2000) uses independence tests on the observed data to infer the appropriate PAG and thus (partial) information on ancestral relationships between the observed variables. We describe the relevant details of the algorithm, along with the necessary adaptations to the time series case in Section 4. For the moment, it suf-

fices to say that the algorithm starts from the complete graph over all variables, and performs a series of independence tests according to which it removes edges between those pairs of variables which are (conditionally) independent. Subsequent 'orientation rules' are employed to derive causal conclusions which, under the stated assumptions and in the limit of correctly identified dependencies and independencies, are guaranteed to be valid. The full algorithm, including a proof that the method not only is sound but that it also is complete, is described in (Zhang, 2008).

3 Time series model

Let $\boldsymbol{X}(t) = (X_1(t), \ldots, X_M(t))$ be a multivariate time series with M variables defined at discrete time points $t \in \mathbb{Z}$, with $\boldsymbol{X}(t)$ either continuous ($\in \mathbb{R}^M$) or discrete-valued ($\in \mathbb{Z}^M$). We assume that the time series data is generated by the following process (essentially a 'dynamic bayes network' with hidden variables, or more precisely a discrete-time first-order time-invariant multivariate Markov process with sparse connections and hidden variables):

1. The causal structure of the time series is defined by a bipartite graph $\mathcal{G}$ with directed edges $\boldsymbol{E} \subseteq \boldsymbol{V} \times \boldsymbol{V}$ where $\boldsymbol{V} = \{1, \ldots, M\}$ is the variable set. An edge $(i, j) \in \boldsymbol{E}$ if and only if variable i has a direct causal effect on variable j.

2. For all j and t, the value of $X_j(t)$ is drawn from a distribution $P(X_j(t) \mid \boldsymbol{X}_{\pi(j)}(t-1)) > 0$ where $\pi(j)$ is the set of parents of variable j.[1]

3. Assume that the process has a strictly positive invariant distribution (or density). For discrete variables with a finite state space, this is guaranteed by the positivity of the conditional distribution, whereas for linear-Gaussian systems the absolute values of all eigenvalues of the co-efficient matrix need to be smaller than unity.

4. The observed data is a subset of size $N \leq M$ of the generating variables. Without loss of generality, let these be the first N variables, i.e. $X_1(t), \ldots, X_N(t)$. The data is observed for time indices $t = \{1, \ldots, T\}$, and the process is

---

[1]Note that for continuous variables, the conditional distributions are typically represented by conditional *densities*.
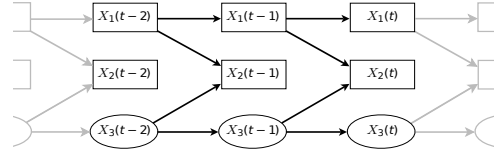

Figure 1: Data generating process over variables $\boldsymbol{X}(t) = (X_1(t), X_2(t), X_3(t))$. The squared nodes represent observed variables, the oval ones hidden.

assumed already to be at equilibrium at time $t = 1$ (i.e. we are sampling from equilibrium).

An example model is shown in Figure 1. In the linear-Gaussian case this model would be represented by $\boldsymbol{X}(t) = \boldsymbol{A}\boldsymbol{X}(t-1) + \varepsilon(t)$, where $\boldsymbol{A}$ is the coefficient matrix and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ represents Gaussian noise. Note that $\boldsymbol{A}$ only has non-zero entries where the corresponding variables are connected by an edge. In the discrete case the parameters would be conditional probabilities; for binary variables there is one parameter for each variable $X_i(t)$ and each (joint) state of the parents of that variable $X_{\pi(i)}(t-1)$.

The model is quite general: For example, higher-order Markov processes can be represented by transforming them to first-order with additional hidden variables, and in terms of linear systems the model can represent any finite-order ARMA model.

Note that in this model the equilibrium distribution of the time series data over any finite-length time window of length τ, i.e. $P(\boldsymbol{X}(t-\tau), \ldots, \boldsymbol{X}(t))$ is well-defined, and we can obtain (correlated) samples from it by taking windowed samples of the time series data. Furthermore, the nodes corresponding to variable-timepoint pairs satisfy the Markov assumption with regard to the complete unrolled graph (although note that dependencies due to nodes 'outside the window' need to be taken into account). Finally, to be able to utilize the constraint-based causal inference framework, we need to require that the distribution $P(\boldsymbol{X}(t-\tau), \ldots, \boldsymbol{X}(t))$ is faithful to the unrolled graph; that is, it cannot contain independencies that are not represented in the structure of the unrolled graph.

4 FCI for time series

Given a single long sequence $\boldsymbol{X}(t)$, $t = 1, \ldots, T$ of multivariate time series data from a causal time

series model with potentially hidden variables, as defined in Section 3, we can obtain correct (but typically only partial) knowledge about causal relationships in the large sample limit as follows:

First, using the 'sliding window' approach, we transform the original time series data into a set of samples of the random vector X which collects the values of all observed variables within a time-window of finite length τ, i.e. $X = (X_1(t - \tau), \ldots, X_N(t - \tau), \ldots, X_1(t), \ldots, X_N(t))^T$. Note that this random vector is of length $(\tau + 1)N$, and we obtain a total of $T - \tau$ samples of it from the data. Since by assumption the observed data come from the time series at equilibrium, each sample is drawn from the equilibrium distribution $P(X)$. (Of course, samples coming from close-by timepoints will be correlated, as in any Markov chain, but as $T \to \infty$ the number of effectively independent samples grows without bound.)

Next, considering each component of X as a separate random variable, the FCI algorithm, designed for non-temporal data, is directly applicable. Given successful independence tests (achievable in the large sample limit) we obtain partial but correct causal inferences concerning the elements of the random vector X, and hence concerning the original time series.

However, the amount of information returned by standard FCI, even in the large sample limit, can be quite restricted (for an example see Section 5). Fortunately, because we know the data came from a time series process, we have much prior information that we can leverage. In particular, we know that (a) causal effects must go forward in time, and (b) because of time-invariance, if $X_i(t - t_1)$ is an ancestor of $X_j(t-t_2)$, then $X_i(t-t_3)$ must also be an ancestor of $X_j(t-t_4)$ whenever $t_1 - t_2 = t_3 - t_4$. By adapting FCI to explicitly incorporate such background knowledge, we not only are able to make more inferences about the existence or non-existence of causal connections, but those inferences we do make are also more often correct, as the prior knowledge regularizes the problem.

Towards this end, we define an edge between $X_i(t-t_1)$ and $X_j(t-t_2)$ to be *homologous* to an edge between $X_k(t - t_3)$ and $X_l(t - t_4)$ if $i = k$, $j = l$, and $t_1 - t_2 = t_3 - t_4$. Because of the time-invariant structure of the data-generating model we obtain the following two Lemmas.

Lemma 1. *On a finite window of length τ including the variables $X(t - \tau), \ldots, X(t)$, if in $P(X)$ we have $X_i(t-t_1) \perp\!\!\!\perp X_j(t-t_2) \mid \{X_k(t-t_3), \ldots, X_l(t-t_K)\}$ with $0 \leq t_i \leq \tau$, then we also have $X_i(t - t_1 + t') \perp\!\!\!\perp X_j(t - t_2 + t') \mid \{X_k(t - t_3 + t'), \ldots, X_l(t - t_K + t')\}$ for all t' such that $(\max(t_i) - \tau) \leq t' \leq \min t_i$.*
Proof. Since the distribution is time invariant, on an infinitely long window the claim is true. When only including τ lags, the independencies are guaranteed to hold if all involved nodes (in particular including those in the conditioning set) lie inside the window. This is ensured by the restrictions on t'. □

Lemma 2. *Taking into account the time-invariant underlying structure, in the PAG corresponding to $X = (X_1(t - \tau), \ldots, X_N(t))$ any two homologous edges must contain the same edge marks.*
Proof. This follows directly from the underlying time-invariant structure, because $X_i(t - t_1)$ is an ancestor of $X_j(t - t_2)$ if and only if $X_i(t - t_3)$ is an ancestor of $X_j(t - t_4)$ when $t_1 - t_2 = t_3 - t_4$. □

The *tsFCI* (for 'time series FCI') algorithm is thus obtained by adapting the FCI algorithm using the prior knowledge, as shown in Algorithm 1. The changes to FCI, highlighted in dark-blue italic, follow from Lemmas 1 and 2. In addition, as suggested in Spirtes et al. (2000), we can in all independence tests restrict ourselves to conditioning sets containing no nodes from the present or the future and orient all arrows forward in time. We emphasize that this prior knowledge needs to be applied at all stages of the algorithm; simply post-processing the output of standard FCI would be suboptimal.

Note that Lemma 1 implies that if a (conditional) independence is found between two nodes $X_i(t - t_1)$ and $X_j(t - t_2)$, this not only allows us to remove that edge from the PAG, but it additionally implies that also some homologous edges can be removed (including all such *later* edges). The directionality is crucial here: earlier edges cannot necessarily be removed, because at the early end of the window we may not be able to condition on the necessary conditioning set. This means that in the PAG, there may be *extra edges* showing up in the early part of the time window. If the time window is short these will allow us to make *fewer* inferences but, in the limit, will not cause us to make *incorrect* inferences.

Algorithm 1 tsFCI (sketch).

Note: Adaptation of FCI algorithm, see (Spirtes et al., 2000; Zhang, 2008) for details of the original algorithm.

Input: A multivariate time series dataset with N variables, an integer τ defining the window length.

1. Adjacency Phase

(a) Get the fully connected graph G over all $(\tau + 1)N$ nodes.

(b) Let Adj_X be the adjacent nodes of a node X in the graph G. Note that G will be modified in the following loop.

    Repeat for $n = 0, 1, 2, \ldots$ (size of conditioning set)

      Repeat for every ordered pair (X, Y) of adjacent variables with $\text{Adj}_{X,past} := |\text{Adj}_X - \{Y\}- \{Z : Z \text{ occurs after or in same time slice as } X\}| \geq n$

        Repeat for every set $Z \subseteq \text{Adj}_{X,past}$ with n elements:

          If $X \perp\!\!\!\perp Y \mid Z$:

            Remove the edge between X and Y in G, define $\text{SepSet}_{X,Y} = Z$
            Remove every homologous edge between $\tilde{X}$ and $\tilde{Y}$ if the corresponding conditioning set $\tilde{Z}$ is fully contained in the graph (by Lemma 1), and define $\text{SepSet}_{\tilde{X},\tilde{Y}} = \tilde{Z}$
            Continue with the next pair.

(c) Try to remove more edges as in FCI by an additional procedure *and use Lemma 1 as in the previous step*.

2. Orientation Phase

(a) *For every edge orient the endpoint in the later time as an arrowhead and for every instantaneous edge orient both ends as arrowheads, by background knowledge.*

(b) Use the complete orientation rule set from Zhang (2008) to orient edge endpoints. *Whenever an edge endpoint is oriented, also orient the same endpoint for every homologous edge (by Lemma 2).* (Note that in this step the "SepSet" from 1(b) are needed.)

Output: A PAG over $(\tau + 1)N$ nodes with repeating structure.

---

5 Simulations

We provide a complete implementation of our algorithm, including code for all of our experiments, at: `http://cs.helsinki.fi/u/entner/tsfci/`

First, to illustrate the theoretical potential (infinite sample size limit) of tsFCI as compared to both Granger causality and standard FCI, consider the example introduced in Figure 1 with window length $\tau = 2$. A Granger-causality approach would correctly infer that the only cause of $X_1(t)$ is $X_1(t-1)$, as this is necessary and sufficient for optimal prediction. However, it would also suggest that there are direct causal connections from *both* variables and *all* lags to $X_2(t)$, due to the hidden variable X_3. FCI and tsFCI yield the PAGs of Figure 2(a) and (b), re-

spectively. Both find that $X_2(t - 1)$ and $X_2(t - 2)$ are not ancestors of (i.e. do not have a causal effect on) $X_2(t)$, and $X_2(t - 1)$ is not an ancestor of $X_1(t)$. From tsFCI (only), we additionally see that $X_1(t-1)$ *is* a cause of $X_1(t)$ and of $X_2(t)$ (FCI could here not rule out a hidden common cause as an explanation for the correlation). One might think that a simple 'post-processing' of the output of FCI would be sufficient but, in practice with finite sample data, this is not the case as the output of FCI may not respect the regular time series structure (see Figure 2(c)).

For a comprehensive analysis, we generate data from randomly constructed time series models, and apply FCI, tsFCI, Granger, Group Lasso (GL) for time series (Haufe et al., 2010), and the 'Phase Slope Index' (PSI) method of Nolte et al. (2008) to the data. Both binary and linear-Gaussian data was used; however we did not have implementations of Granger or GL for binary data, and PSI is only applicable to continuous data. We can deduce the behavior of FCI, tsFCI, and Granger in the infinite sample limit by using an independence oracle based on d-separation in the generating model.

The methods are evaluated based on three scores, each measuring the percentages of correct vs incorrect vs "don't know" answers to a set of simple causal queries. The 'direct-cause' score asks if a given node (variable-timepoint pair) has a direct effect on another given node, the 'ancestor' score asks if a given node has any effect (potentially indirect) on another node, and the 'pairwise' score asks if a given variable has an effect on another given variable (with any lag). Note that GL and PSI are specifically constructed to answer the last question.

For reasons of space we cannot give all details of the data generation here, but they are thoroughly described in our online code package. Briefly, we used models with up to 6 observed and 3 hidden time series variables, with edge probabilities $q = 0.25$ and $q = 0.5$ between any pair of nodes, and generated data of three different sample sizes (100, 1,000, 10,000). The window length is fixed at $\tau = 3$ and we set the significance level to 0.01.[2] Results are shown in Figure 3. We state a few main points:

First, in the infinite sample size limit ("$T \to \infty$"),

---
[2]Simulations showed that the significance level seems not to have a crucial effect on the results. For smaller window lengths τ less decisions are made than for longer ones.
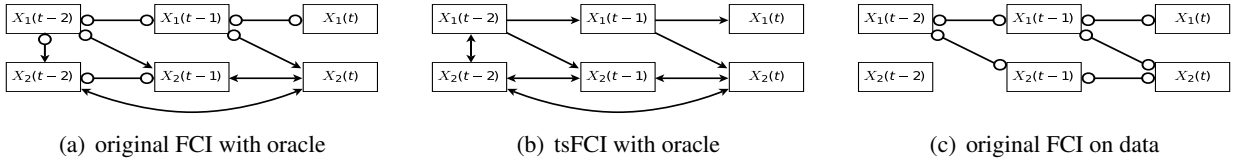
Figure 2: Output PAG of FCI and tsFCI when applying the algorithms to the generative model in Figure 1, in (a) and (b) with an oracle (infinite sample size limit), and in (c) for a small sample size.

the main difference between the FCI-based methods and Granger causality is that with the former some decisions are not made but all made decisions are correct, whereas in the latter all decisions are made but some of them are wrong (because of the latents). In this sense FCI/tsFCI is (in theory) more conservative than Granger. As expected, tsFCI makes somewhat more predictions than FCI.

Second, tsFCI consistently outperforms FCI on finite-length data as well, particularly for continuous variables and the ancestor score (structural mistakes of FCI may be reflected in this case).

Third, on all scores and for all methods, the accuracy increases with sample size, as expected. However, for realistic sample sizes, FCI and tsFCI are quite far from the infinite sample limit. In particular, we found that as the unrolled graphs were relatively dense (even for $q = 0.25$), these methods need to rely on independence tests with large conditioning sets, leading to inevitable errors that accumulate. Only for the sparse, binary case does the performance approach the theoretical limit. For the continuous data, with our test settings, the standard Granger procedure is superior to both FCI-based methods both in terms of number of decisions made *and* in terms of the correctness of those decisions.

Fourth, for binary data, the number of decisions *decrease* for larger samples. This counterintuitive result is due to weak causal effects (over long paths) detected only for these sample sizes, which yields to more edges and hence fewer orientation rules might be applied. This is not particular to our case, but can occur in other applications of FCI (to DAG inference) as well.

Finally, in our experiments neither PSI nor Group Lasso outperformed a basic Granger analysis. Since we tried to choose the parameters close to optimal for these methods, the explanation for our results is that the particularities of the data generating setup

favored Granger. For instance, PSI was at a clear disadvantage here because it seeks driver-receiver-relationships among the variables, and so does not allow both variables in an interacting pair being drivers of the other. Such relationships were however common in our data.

6 Firm growth data

We also applied the tsFCI algorithm to microeconomic data of growth rates of US manufacturing firms in terms of employment, sales, research & development (R&D) expenditure, and operating income, for the years 1973–2004. Here, we estimated the model including instantaneous effects,[3] using a significance level of 0.001 and $\tau = 3$. The first two lags of the PAG are shown in Figure 4.

The graph suggests that there are direct causal effects over time from employment to sales growth and from sales to employment growth. Furthermore, there is no causal effect from operating income to R&D growth (instantaneously or lagged), whereas both employment and sales growth have (direct or indirect) lagged effects on R&D growth. Interestingly, these features essentially duplicate the main findings of Moneta et al. (2010), in which a Structural Vector Autoregression model was used to infer causal relationships among the variables. One difference, however, is that their analysis suggested significant contemporaneous causal effects among the variables, while our present approach attributes most of the correlations to latents.

7 Related work

Most standard approaches to causal inference from non-experimental time series data are rooted in the concept of Granger causality (Granger, 1969). In

---

[3]In Algorithm 1 this means to not exclude nodes from the present in the conditioning sets in step 1(b) and to not orient instantaneous edges as double-headed arrows in step 2(a).
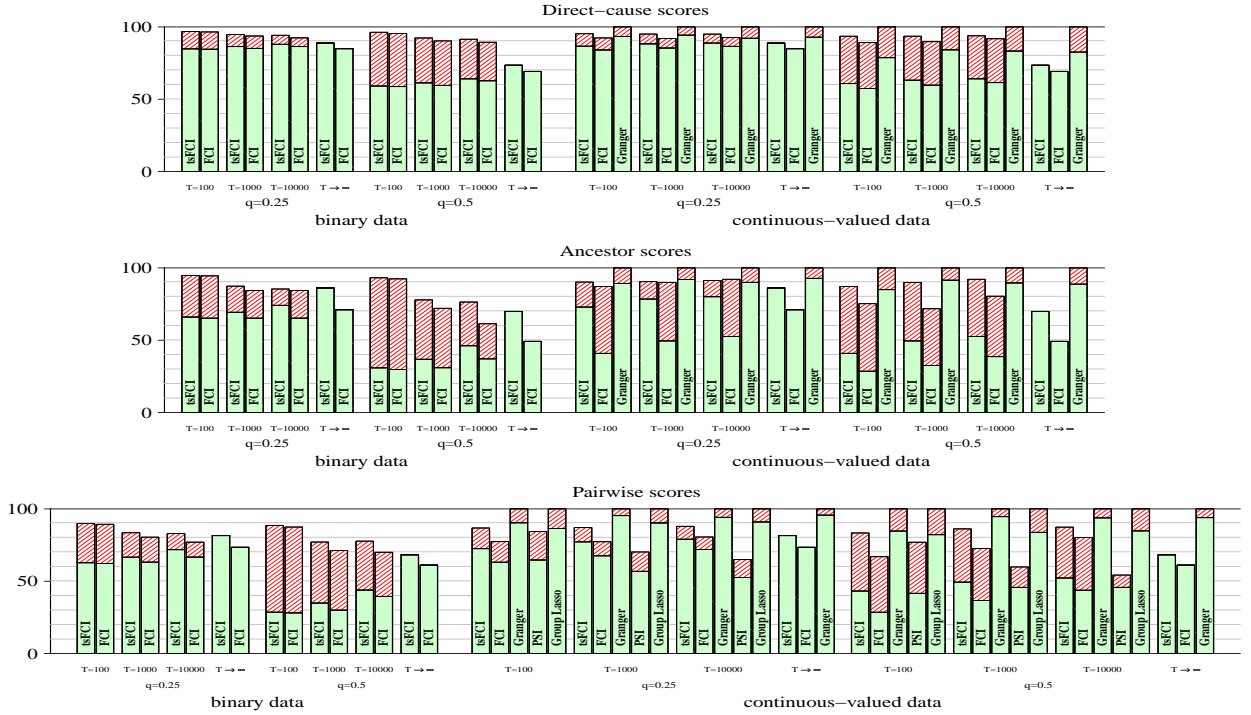
Figure 3: Results of simulations. From top to bottom row: direct-cause score, ancestor score, pairwise score. In each row, the two left blocks of bars show the result for binary data for sparse (q=0.25) and less sparse (q=0.5) graphs for different sample sizes and algorithms. The two right blocks of bars represent the same results for continuous-valued data. The total height of each bar shows the percentage of made decisions. Within each bar the green filled area marks the correct decisions, the red striped area the incorrect decisions.

essence, the assumption is that all relevant variables have been measured, and there are no contemporaneous effects, in which case the problem of determining causality turns into a well-defined statistical estimation problem. Naturally, it is important to note that the chosen algorithms and representations (such as whether to use a time-domain or frequency-domain representation) are important in determining the overall performance of the algorithm from a finite-length time series; see, for instance, the work of Nolte et al. (2008) and Haufe et al. (2010).

One departure from the standard assumptions is to include contemporaneous causation. For linear models, this amounts to using Structural Vector Autoregression (SVAR) models rather than the Vector Autoregression (VAR) model. The extra causal information inherent in SVAR models can be inferred using conditional independence tests (Swanson and Granger, 1997; Demiralp and Hoover, 2003) or utilizing non-Gaussianity (Hyvärinen et al., 2008). Nonlinear additive models, with latents which are

uncorrelated over time (Chu and Glymour, 2008), constitute an interesting extension of this line of work. However, these models are not guaranteed in the limit to give correct results when temporally dependent hidden variables may be present.

Another interesting recent development is given by the *difference-based causal models* of Voortman et al. (2010), specifically designed to model systems that can be well represented by difference equations. For such systems, their framework is likely to be superior to more general models in terms of inferring the causal structure.

Finally, the strongest connection to other recent work is to the theoretical work of Eichler (2009), for which, unfortunately, at the time of writing no detailed algorithm was published to compare our method against. In his approach, the starting point is the estimation of Granger-causal vs Granger-non-causal relationships, and any remaining contemporaneous dependencies. From such a *path diagram*, certain causal vs non-causal inferences can be made
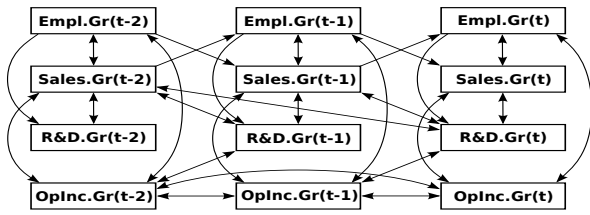
Figure 4: Partial ancestral graph inferred from the firm growth data (see Section 6 for discussion).

which are robust with respect to the presence of hidden variables. His method is also, though much less directly than ours, derived from the theory developed for DAG structures. The main difference between the two approaches is that tsFCI explicitly models the temporal dynamics. This has the advantage of more often being able to detect the absence of direct effects between variable pairs. This comes at a cost, however: For short time series it may be difficult to reliably infer all temporal dependencies, in which case the predictions may be less reliable.

8 Conclusions

While the assumptions underlying FCI seem in many respects more likely to hold than those needed to obtain valid causal conclusions from a Granger-causality approach, our study suggests that caution is in order. Even for processes with relatively few interacting components, the practical problem of detecting which independencies hold (and which do not), from time series of reasonable length, can be a significant obstacle to reliable causal inference. Nevertheless, we suggest that, when the results are interpreted with due caution, tsFCI may help shed light on causal interactions in time series data.

Finally, while the models we consider are very general, there is one respect in which they are quite restricted: We cannot model cyclic contemporaneous causation. In many real world scenarios one may not have a high enough sampling rate to ensure that such causation within a measurement period does not exist, including feedback loops. Unfortunately, such loops are a violation of a basic assumption underlying FCI. Furthermore, many time series are aggregates in the sense that each data point is a sum or average of a number of time slices in an original time series with much faster dynamics. In such cases the independencies present in the measured data do not necessarily reflect the causal structure in the original data. A challenge is how to extend present methods to handle this type of data.

Acknowledgements

References

Chu, T. and Glymour, C. (2008). Search for additive nonlinear time series causal models. *JMLR*, 9:967–991.

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

Eichler, M. (2009). Causal inference from multivariate time series: What can be learned from Granger causality. In *13th International Congress on Logic, Methodology and Philosophy of Science*.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438.

Haufe, S., Müller, K.-R., Nolte, G., and Krämer, N. (2010). Sparse causal discovery in multivariate time series. *JMLR W&CP*, 6:97–106.

Hyvärinen, A., Shimizu, S., and Hoyer, P. O. (2008). Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *Proc. ICML*, pages 424–431.

Moneta, A., Entner, D., Hoyer, P. O., and Coad, A. (2010). Causal inference by independent component analysis with applications to micro- and macroeconomic data. *Jena Economic Research Papers*, pages 2010–2031.

Nolte, G., Ziehe, A., Nikulin, V. V., Schlögl, A., Krämer, N., Brismar, T., and Müller, K.-R. (2008). Robustly estimating the flow direction of information in complex physical systems. *Physical Review Letters*, 100:234101.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

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

Richardson, T. S. and Spirtes, P. (2002). Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030.

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

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

Voortman, M., Dash, D., and Druzdzel, M. J. (2010). Learning why things change: the difference-based causality learner. In *Proc. UAI*.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172:1873–1896.