Using Domain Knowledge to Overcome Latent Variables in Causal Inference from Time Series

Min Zheng MZHENG3@STEVENS.EDU

Samantha Kleinberg SAMANTHA.KLEINBERG@STEVENS.EDU

Computer Science
Stevens Institute of Technology
Hoboken, NJ, USA

Abstract

Increasingly large observational datasets from healthcare and social media may allow new types of causal inference. However, these data are often missing key variables, increasing the chance of finding spurious causal relationships due to confounding. While methods exist for causal inference with latent variables in static cases, temporal relationships are more challenging, as varying time lags make latent causes more difficult to uncover and approaches often have significantly higher computational complexity. To address this, we make the key observation that while a variable may be latent in one dataset, it may be observed in another, or we may have domain knowledge about its effects. We propose a computationally efficient method that overcomes latent variables by using prior knowledge to reconstruct data for unobserved variables, while remaining robust to cases when the knowledge is wrong or does not apply. On simulated data, our approach outperforms the state of the art with a lower false discovery rate for causal inference. On real-world data from individuals with Type 1 diabetes, we show that our approach can discover causal relationships involving unmeasured meals and exercise.

1. Introduction

Causal relationships are why we can successfully predict future events like illness, intervene to change outcomes such as by reducing risk, and explain why events like a particular person's illness happened. Health data such as from electronic medical records, intensive care unit data streams, and patient generated health data are becoming more widespread and could potentially be used to uncover causes of illness. However, these observational datasets were not collected for research and, critically, we rarely have control over which variables are measured. This violates a core assumption of many causal inference methods: no latent common causes. When a shared cause of two or more variables is absent, we may find spurious relationships between the variable's effects. This problem is particularly important in health, as the result of our inferences may inform treatments. False inferences could lead to ineffective interventions to treat disease that potentially put patients at increased risk.

Due to the fundamental nature of this problem, many solutions have been proposed, and most augment methods for structure learning in Bayesian networks (Pearl, 2000). However, these methods do not address the more challenging problem of latent causes in temporal data. Methods for time series such as tsFCI (Entner and Hoyer, 2010) bring increased computational complexity and have not yet been successful with large datasets with many densely connected variables. One way to improve both computational complexity and accuracy is by leveraging prior knowledge, such as information about conditional independence relationships (Hyttinen et al., 2013). Such prior information has not been leveraged with temporal relationships, though, where we must also know the timing of the relationships – bringing further chances for errors.

Yet, even if a variable is latent in one dataset, it may not be latent in every dataset. Variables may be latent for many reasons, including cost and difficulty of measuring them (e.g. invasive test), context (e.g. hospitals have different monitoring protocols), and time (e.g. new measurement technology developed). Further, we may also have an understanding of the expected effects of a variable based on domain knowledge. However, this knowledge may be incorrect, or may not apply to the dataset at hand. For example, a variable may be measured in a population with a particular type of health insurance, who may have different risk factors. Thus it is critical that we can identify when prior information does not apply. To address this, we propose a new method for causal inference in time series with latent variables that leverages prior knowledge (in the form of causal relationships), discovers whether it is inconsistent with the data, and uses the applicable information to reconstruct when latent variables may have occurred. We show that this can be used to identify latent variables and infer causes with significantly higher accuracy than tsFCI (Entner and Hoyer, 2010) on simulated datasets. We apply the approach to real-world data from individuals with Type 1 diabetes (T1D), demonstrating that our approach can be used to accurately infer causes and effects of meals, even when this information is latent.

Technical Significance Causal inference is a core problem in machine learning, but the strong assumptions of most methods have led to limited applicability in real-world healthcare data. In particular, many methods for learning causal structures from data assume no latent variables, but this is clearly violated in real-world data. While some methods exist for inference with latent variables in time series data, they have low accuracy in realistic settings, and methods that incorporate prior knowledge assume this information is correct. In contrast, we propose a novel method that exploits prior knowledge to overcome latent variables, and crucially does not rely on the correctness of this prior knowledge. We show experimentally that our proposed approach outperforms the state of the art on simulated data, and can successfully leverage prior information with real-world data.

Clinical Relevance Large amounts of medical data are being collected both in hospitals and by patients in daily life, but the incomplete nature of these observational data makes it challenging to apply cutting edge machine learning methods, which often make strong assumptions that do not hold in health data. Our clinical contributions are two-fold. First, identifying causes rather than correlations is critical to making effective treatment decisions and ensuring that interventions are not simply addressing a symptom but rather the underlying issue. Our approach is broadly applicable to many health time series. Second, we apply our methods to better understand T1D, which affects 9% of the population worldwide. While data from daily life is important for identifying factors affecting blood glucose (BG), it is difficult to collect data on meals and exercise longterm. We show how general knowledge about T1D can be used to overcome cases where these variables are missing, allowing better use of these data to gain insight into causes of changes in BG.

2. Related Work

Most related work on causal inference with latent variables is based on Bayesian networks (BNs) (Pearl, 2000), which use directed acyclic graphs to represent causal relationships, and probabilistic independence relationships and a set of assumptions to infer structure from data. One such approach that can handle latent variables is Fast Causal Inference (FCI) (Spirtes et al., 2000), which first finds partial ancestral graphs (PAGs), then orients their edges (Zhang, 2008). FCI has exponential time complexity, which limits its applicability to large data. There have been a number of extensions of FCI to improve both its computational efficiency and applicability: Really Fast Causal Inference (RFCI) (Colombo et al., 2012) reduces the complexity of FCI by using fewer independence tests; Anytime FCI (Spirtes, 2001) terminates conditional independences test early to reduce the search space; and FCI-Max (Raghu et al., 2018) and FCI-Stable (Colombo and Maathuis, 2014) allow a mix of continuous and discrete datasets. Other approaches constrain the type of BNs that can be inferred. For example, Zhang (2004) proposed a method to learn hierarchical latent class models, but this only covers BNs where the structure is a tree and all nodes are latent except the leaves. Silva et al. (2006) also constrains the structure so that observed variables are leaves of a tree but with strong assumptions that unmeasured (latent) variables cannot be effects of observed variables and that dependencies are linear. However, none of these approaches are applicable to time series data.

Methods that address latent variables (also referred to as latent confounders, or hidden confounders) in time series often build on both BNs and FCI. BNs have been extended to include time using Dynamic Bayesian networks (DBNs) (Murphy, 2002), and extensions have been developed to handle latent variables in DBNs. For example, Song et al. (2009) proposed Time-Varying DBNs to recover latent networks underlying biological processes. However, the structures are required to be sparse and to vary smoothly across time. FCI has been extended to time series with tsFCI (Entner and Hoyer, 2010), which transforms the time series into random variables, then applies FCI to these variables. Another FCI-based method combines Granger causality (Granger, 1980) (a method for causal inference in time series) and FCI, by representing the coefficient matrix with a path diagram before applying FCI (Eichler, 2010). However, all methods that use FCI share its high computational complexity and since the order of independence tests matters, errors can propagate. Finally, Voortman et al. (2010) learns difference-based causal models from time series to avoid latent confounders, but data are assumed to be generated by differential equations.

Other methods make use of information beyond the data to deal with latent variables. Hyttinen et al. (2013) incorporate information about conditional independences to learn cyclic causal structures. While this method allows feedback loops and does not have parametric restrictions (e.g. linearity), it relies on the correctness of the conditional independence statements. However, even if such a statement is correct in one setting, it may not apply for example to a different population. Borboudakis and Tsamardinos (2012) proposed a method to handle latent variables using knowledge of existence or non-existence of causal relationships. That method aims to find structures that are consistent with knowledge, while considering the degree of belief in each piece of prior knowledge. Since beliefs may be incoherent, this approach was extended to to handle dependent and incoherent beliefs, however this comes at the expense of exponential time complexity (Borboudakis

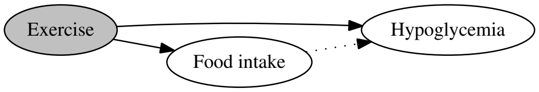


Figure 1: Structure showing effect of latent variables (shaded in grey). The hidden common cause may lead to a spurious relationship between its effects (dotted edge).

and Tsamardinos, 2013). Cooper and Herskovits (1992) used prior knowledge of the temporal ordering of all variables to handle both missing data and latent variables by using a greedy search strategy (K2) when inferring a BN. Chen et al. (2008) improved K2 by using independence tests to reduce the search space. However, both approaches rely on the the variable ordering, and cannot overcome errors in prior knowledge. Further, they have not yet been applied to time series data. In contrast, we use prior knowledge (that may be incorrect) to reduce the search space, and allow that latent and observed relationships can have arbitrary time lags.

3. Method

We introduce a new approach to causal structure learning for time series with latent variables, using the idea that prior information can be used to augment data, since latent variables are not always latent in all datasets. We begin with an overview of the problem and details on the method we extend, before introducing our approach.

To motivate our approach, consider trying to understand BG in people with T1D. Physical activity, meals, and stress affect BG, but not all are measured in every study, due to the time and expense incurred with each extra variable. Thus we may know that moderate physical activity can lead to hypoglycemia, and that exercise tends to make people eat (e.g. hunger or overcompensating for calorie burn), but we may only be able to measure BG and food intake. In that case we may find a paradoxical relationship where eating seems to lower BG, as shown in Figure 1, because there is a latent common cause of the two observed variables but one regularly happens before the other. This is what makes the temporal case more challenging than latent variables in static data. However, this relationship can also differ between individuals and contexts, such as a competition where adrenaline leads to hyperglycemia even with moderate activity. The question we aim to address is: can we leverage such information when available to avoid confounding, while being robust to small errors in knowledge?

3.1. Background

We are focused on overcoming confounding due to latent variables in causal inference from time series data, while avoiding high computational complexity. We build on the approach of Kleinberg (2012), for exact inference of causal relationships and their timing as it is $O(N^3T)$ (N being the number of variables and T the length of time series). Causal relationships are represented by logical formulas, where each can include arbitrarily complex causes and effects, and the cause brings about the effect in some window of time, such as:

$$m \leadsto^{\geq 15, \leq 30}_{\geq 0.7} h, \tag{1}$$

meaning that moderate exercise m leads to hypoglycemia h in 15 to 30 minutes with probability 0.7.

To infer causal relationships from time series data, the approach tests user generated logical formulas, or relationships between pairs of variables up to a specified level of complexity. With time series data and a set of logical formulas representing causal relationships, the goal is to find which are significant. We focus on the case where all variables are discrete, so potential causes raise the probability of their effects. Then c is a potential cause of e if $P(e|c) > P(e)$. To distinguish between spurious factors and genuine causes, the approach uses a measure of the significance of causes that indicates how much the probability of the effect changes in the presence of the cause, once all other potential causes are held fixed. Causal significance is defined as:

$$\varepsilon_{avg}(c_{r-s}, e) = \frac{\sum_{x \in X \backslash c} P(e|c \wedge x) - P(e|\neg c \wedge x)}{|X \backslash c|}, \tag{2}$$

with relationships of the form $c \leadsto^{\geq r, \leq s}_{\geq p} e$ and $x \leadsto^{\geq r', \leq s'}_{\geq p} e$. $P(e|c \wedge x)$ is the probability of e in time window $[r, s]$ after both c and x occur. Relationships that are statistically significant, where ε is greater than a threshold, may be causal. For these to be guaranteed to be causal, one must assume that there are no hidden common causes and that relationships are stationary over time.

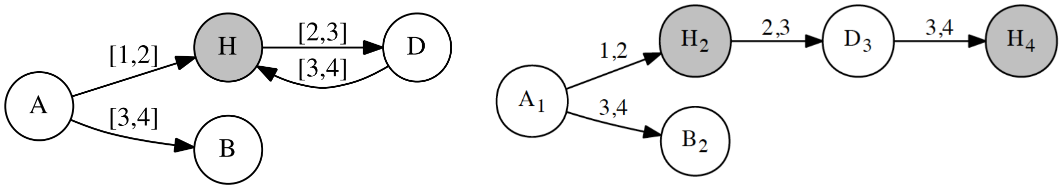
3.2. Preliminaries

We use uppercase letters to denote sets of variables, and lowercase to denote individual variables in the set. With a time series of length T and Boolean-valued variables V, the data can be represented by a $T \times V$ matrix, D, where $D(i, j) = 1$ if v_j is true at t_i.

We assume a *knowledge base* K, which is a set of causal relationships of the form in equation (1). Relationships in K can include variables outside of V, and thus may capture information about latent common causes. K may be inferred from a different dataset or may come from domain knowledge. In some cases D may be an incomplete portion of a dataset, and K relationships learned from the complete portion. We allow that K may be incorrect, and will aim to discover which relationships do not apply to our dataset.

3.2.1. Assumptions

To guarantee correctness, meaning that inferred causes are genuine and we can remove all confounding, we must make the following assumptions. First, the true set of causal relationships underlying time series D for the set of variables V must be *stationary*. That is, causes and their significance do not change over time. Second, the data must be *faithful* to this structure (as in Spirtes et al. (2000)), so the true set of causes is always entailed by D. Third, K must include at least one true relationship that includes each latent common cause of variables in D. Note that this is a weaker assumption than requiring knowledge of the causal relationship between the latent variable and its observed effects. If the latent



(a) Relationships in K

(b) Unrolled DBN

Figure 2: Illustration of time-windowed graphical model construction.

variable is in the knowledge base, we are able to reconstruct it and will be able to learn about the relationships with observed effects.

3.3. Causal inference with latent variables

Our approach has three parts: 1) using knowledge from K to reconstruct the time series of latent variables in a dataset D; 2) identifying errors in K and updating the reconstructed data; 3) inferring causal structures from the observed and reconstructed dataset.

3.3.1. Reconstruct time series for latent variables

We now discuss how to use K to reconstruct when each latent variable ($v \notin V$) occurred. We refer to the observed data as D. The output of this step is D', which augments D with the recovered time series. As shown in Figure 1, when exercise is latent but we observe BG and meal times and know how exercise affects both, we can use observations of BG and meals to identify when exercise was likely to have occurred. Not every instance of a meal is preceded by exercise, and in general a latent variable may have multiple observed causes and effects that can be used as evidence. Thus we aim to infer the probability of each latent variable at each time – rather than assuming it must deterministically precede each effect.

To do this we build on DBNs, with two key updates. First, since exact inference is NP-hard, when inferring the probability of v, we use only variables in its Markov blanket (parents, children, and children's parents). This significantly reduces the number of evidence variables needed during inference. Second, while DBNs allow temporal relationships they use discrete lags (e.g. X causing Y in 10 time units, 11 time units, 12 time units and so on), while we build on methods that use time windows (X causes Y in 10–20 time units). Thus we augment DBNs with time windows, which further improves efficiency by allowing multiple edges to be collapsed into one, meaning that a probability is inferred once for the range rather than for each individual lag. We further improve efficiency by caching results to avoid redundant computations as we sequentially impute values over time.

We begin by building a DBN, G, using the relationships in K. For each relationship $(c \leadsto^{\geq r, \leq s}_{\geq p} e) \in K$ we add an edge $c \rightarrow e$ to G, but instead of the edge being to e at a single future time lag t, it is parameterized with the time window $[r, s]$. Figure 2(b) shows the unrolled graph, when K is $\{A \rightarrow B, A \rightarrow H, H \rightarrow D, D \rightarrow H\}$. We do not allow instantaneous relationships, but feedback loops that happen across time can be handled, just as in DBNs.

For each latent variable $v \notin V$, we aim to determine when it may have occurred. To do this, we infer $P(v|MB(v))$ at each time $t \in T$, where $MB(v)$ is the Markov blanket of v. The Markov blanket of v is defined by:

$$\mathrm{MB}(v) = \mathrm{pa}(v) \cup \mathrm{ch}(v) \cup \bigcup_{y \in \mathrm{ch}(v)} \mathrm{pa}(y) \tag{3}$$

where $\mathrm{pa}(v)$ denotes parents of v and $\mathrm{ch}(v)$ denotes children. In Figure 2(*b*), $MB(H)$ includes all variables except H. $MB(v)$ is comprised of both observed variables (E), and latent variables (Y). Observed variables, evidence in our inference, are set to their actual values from D. Then, the probability of v_t given our observations is:

$$P(v_t|MB(v_t)) = \frac{\sum_{y \in Y} P(v_t, Y, E)}{\sum_{y \in Y} P(Y, E)}. \tag{4}$$

The value of evidence variables E, will depend on the time t. Because of our inclusion of time, parents of v will have relevant data before t, while children will have relevant data after. Since the causal relationships have time windows, the specific times used for each variable will vary. Further, there may be multiple observations of a variable. The value of an evidence variable $e \in E$ is set as follows:

$$e = \begin{cases} Max(D[e, t-s] \dots D[e, t-r]), & \text{if } e \in pa(v) \\ Max(D[e, t+r] \dots D[e, t+s]), & \text{if } e \in ch(v) \\ Max(\bigcup_y Max(D[y, t+r] \dots D[y, t+s])) \\ \text{where} \quad y \in (ch(v) \cap ch(e)), \text{and} \quad e \in pa(ch(v)) \end{cases} \tag{5}$$

Thus for parents (causes) of v we test whether they occurred before t in the time window $[r, s]$ associated with the causal relationship, and conversely test later times for children of v. For children of v, we iterate over each parent of each $ch(v)$, testing whether it occurred before each $ch(v)$ within time window $[r, s]$ of the relationship between $pa(ch(v))$ and $ch(v)$.

We incorporate the state of all other observed variables in v's Markov blanket, while accounting for the different time windows of each relationship. Then, we set the value of v_t in the reconstructed time series D':

$$D'(v, t) = \begin{cases} 1 & \text{if } \mathrm{B}\left(P(v_t|MB(v_t))\right) \\ max(0, D(v, t)) & \text{otherwise} \end{cases} \tag{6}$$

where $\mathrm{B}\left(P(v_t|MB(v_t))\right)$ indicates choosing from a Bernoulli distribution. Since there may be multiple instances of the same values in $MB(v_t)$, especially as the length of the time series grows, we store results for each setting of evidence, so these need not be recomputed each time, significantly reducing the computational complexity of the approach. Experimentally, we show how this significantly reduces the number of computations required compared to the theoretical maximum.

3.3.2. Identify errors in K

In the first step we use prior knowledge to learn when latent variables may have occurred. However, since we do not assume this knowledge is completely accurate, it is possible that if we reconstruct data based on K, some of D' may be incorrect. Thus, in this step we aim to identify which relationships in K do not apply. According to our assumptions, the true set of conditional independence statements will always be entailed by D. Therefore, if we find that the conditional independence relations entailed by K are different than those entailed by D, then K has errors. However, since the same set of conditional independences can correspond to multiple causal structures, an inconsistency only indicates that an error exists in K. It does not directly yield the correct causal structure. Thus, once we find inconsistent conditional independencies between D and K, we generate a set of Markov equivalent causal structures to replace K. We use these Markov equivalent structures to again reconstruct D' using the process described above until we find consistent conditional independences between D' and K.

After recovering a latent time series using K, we perform conditional independence tests on the recovered time series D' to get a set of conditional independence statements M' entailed by D'. For each pair of variables $X, Y \in D'$, we iterate over each possible time window $[r, s]$ given each possible subset of variables in D'. For each X, Y, we compute the absolute difference between the following two conditional probabilities given each subset of variables $U \subset D' \setminus \{X, Y\}$ as follows:

$$\phi_{X,Y|U} = 1 - |P_{[r,s]}(X|Y) - P_{[r,s]}(X|Y, U)| \tag{7}$$

We add $X \amalg Y|U$ to M' when ϕ is statistically significant (experimentally we use $p < 0.05$). Let M be the set of conditional independences entailed by prior knowledge K. If $M \neq M'$, it means K has errors. For each pair $X, Y \in D'$ that is conditionally independent only in M or in M', we generate a set of Markov equivalent structures including variables in $\{X, Y\} \cup MB(X) \cup MB(Y)$. We regenerate the time series for each of the Markov equivalent structures using the process described. We restrict ourselves to the Markov blanket of X, Y to minimize the number of variables to be conditioned on, while capturing key relationships. We repeat the time series reconstruction procedure discussed in the previous section until $M = M'$. Since the true causal relationships will always be entailed by the data and the set of Markov equivalent structures generated includes all possible structures, the true structure that yields $M = M'$ will always be found.

3.3.3. Infer causal structures

After augmenting D with the time series of latent variables to obtain D', we then apply the causal inference method of Kleinberg (2012) to D' to learn a set of causal relationships.

3.3.4. Complexity

There are three components of our algorithm: inferring data for the latent variables, identifying errors, and lastly causal inference. To infer the time series for each latent variable v, we compute its probability at each timepoint, using evidence from its MB. In the worst case, the MB includes all variables and all variables are latent. Since inference is done for each timepoint, the worst case complexity is then $O(N^3TN) = O(N^4T)$. However, since we cache results during inference, it is not necessary to recalculate the probability of each element of the MB for each v_t. Therefore, the overall complexity for inferring the latent series is $\ll O(N^4T)$. To identify errors, we need to test each pair of variables, which is

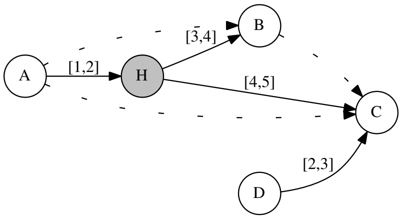


Figure 3: Structure with latent variable H (shaded circle) and spurious relationships (dotted edges).

$O(N^2)$. In the worst case, all relationships in K are wrong and we need to reconstruct the time series for all variables again, which is then $O(N^4T)$. However, as the number of latent variables is smaller than the number of observed variables, and we only reconstruct the variables involved in the Markov blanket, the complexity is $< O(N^4T)$. Finally, causal inference with N variables and time series of length T is $O(N^3T)$ when testing pairwise relationships. Therefore, the overall algorithm complexity is $< O(N^4T)$.

4. Experiments

We first evaluate our method and compare it to the state of the art on simulated data where ground truth is known, demonstrating that our approach can use prior knowledge of latent variables to avoid confounding. Then, on a real-world T1D dataset we show that by including prior knowledge our approach is able to make novel inferences that others cannot.

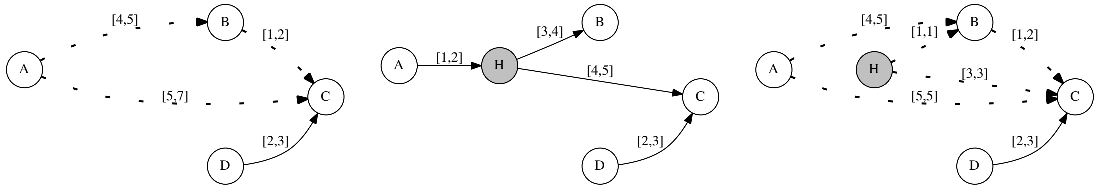
4.1. Simulation

4.1.1. Data

We simulate two types of data: 1) a simple model that has one latent variable that is a common cause of observed variables and part of a chain, and 2) a range of complex models with prior knowledge that varies in completeness and correctness.

Simple We simulate a classic case of confounding where two observed variables have a shared cause that is latent (Figure 3). This case is even more challenging when there are time windows, and we may incorrectly find B as a cause of C when H is latent. The latent variable has an observed cause, A, which may also be inferred as an indirect cause of B and C. The prior knowledge K includes $A \rightarrow H$ and $H \rightarrow B$, and the two causal relationships' time windows and probabilities. We simulate 5000 timepoints for both simple and complex datasets. For simple datasets we add two noise variables (not involved in causal relationships). All time windows are in [1,5] (simple) or [1,6] (complex) but we test [1,8] to make the task more difficult. Causal relationships are simulated with probability 0.9.

Complex To test our approach on more realistic data, we generate datasets with 20 to 100 variables. For each number of variables (incrementing by 20) we generate 10 datasets, for 50 total (5 different variable sizes, 10 datasets for each). The average in/out degree of the simulated structures ranges from 2.5 (20 vars) to 4.5 (100 vars). Causal relationships



(a) Kleinberg (2012) result (b) Our result (c) tsFCI result

Figure 4: Relationships found a) when H is latent, b) after inferring H's time series with our approach, and c) by tsFCI. Solid edges are correct inferences, and dotted are spurious. Edges are annotated with inferred time windows. Ground truth is shown in Figure 3.

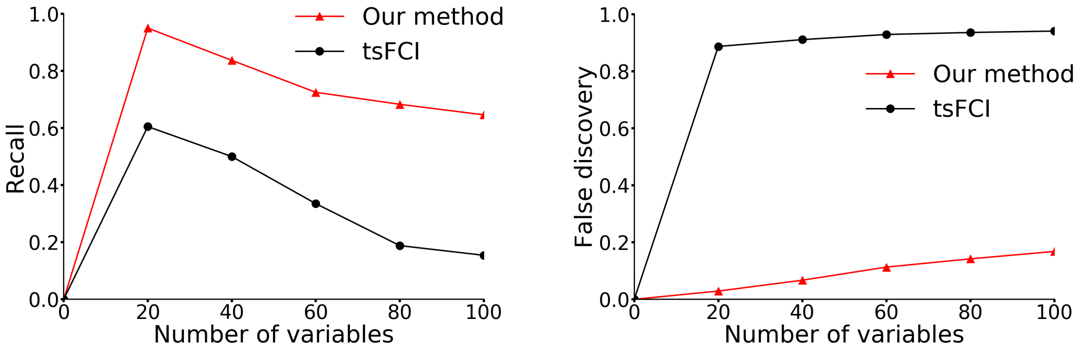
have a randomly generated probability in [0.8,0.9], and we randomly select 20% of variables to be latent in each structure. Prior knowledge, K is generated as follows. We take the set of all causal relationships involving latent variables (L) and 1) randomly remove 10%, 2) after removing relationships we randomly add error to 10% of the remaining ones in either timing (making the window larger or smaller) or probability (increase or decrease), and 3) add 5% of $|L|$ incorrect relationships. The incorrect relationships could be existing latent ones that were not previously perturbed, in which case the arrow is reversed (e.g. if ground truth is $A \rightarrow B$, K will have $B \rightarrow A$), or may be randomly generated new (incorrect) causal relationships. Thus if L has 40 remaining relationships after step 1, K will ultimately include two totally incorrect causal relationships.

4.1.2. Methods

For our approach, we determine which causal relationships inferred are significant using $p < 0.01$ due to the large number of comparisons. We compare our method with tsFCI using the following settings. We test tsFCI using the RCode_TETRADjar package with inclIE=true (no instantaneous effects). We set the parameter nrep (number of time lags plus one) based on the ground truth of the data. The evaluation metrics are recall (what fraction of latent variables are recovered) and false discovery rate (what fraction of inferred causal relationships are false). A latent variable v is correctly recovered if at least one correct causal relationship involving v is inferred. To be correct, inferred causal relationships must also have the correct time window.

4.1.3. Results

Simple structure Results are shown in Figure 4. We first applied the method of (Kleinberg, 2012) (Figure 4(a)), which finds the indirect relationships between A and H's children, as well as an incorrect relationship between B and C, since a core assumption of the method is violated. As expected, when the common cause is latent and one effect regularly occurs before the other, a causal relationship between them is inferred. Similarly, tsFCI finds both the indirect relationships, and the confounded one between B and C. It also identifies the true relationships between the latent variable and its effects, but identifies the wrong timing for them. Further, while this structure has only a single latent cause, tsFCI returns



(a) Recall of latent variables (b) FDR of causal relationships

Figure 5: Comparison of our approach and tsFCI on simulated complex structures.

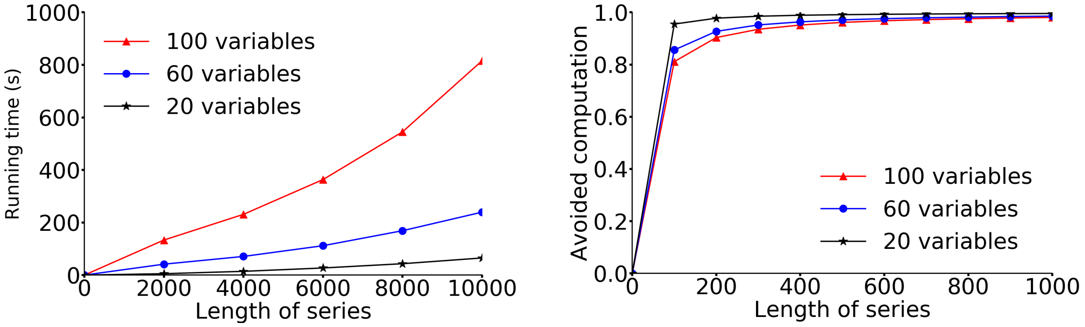
11 latent variables. In contrast, our approach finds only the direct relationships at their correct times.

Complex structures Figure 5(a) shows that our approach correctly recovers more latent variables than tsFCI. Most importantly, as shown in Figure 5(b), our approach has a significantly lower FDR than tsFCI for inference of the causal structure. When the causal structure has 20 variables, our recall of latent variables is 95.0% with a causal inference FDR of 2.9%. However, for the same case, tsFCI has lower recall (60.5%) and a significantly higher FDR of 88.7% (Figure 5(b)). As structures become more complex, more instances of each latent variable may be inferred, making its time series less informative. While we see that as the structures become more complex (and the number of erroneous relationships in prior knowledge increases), the recall of latent variables does decrease, we still recover significantly more than tsFCI and further, the accuracy for causal inference remains significantly higher. Going from 20 variables to 100 variables, tsFCI recall of latent variables drops 45.1%, while ours drops 30.4% (to 66.4% compared to 15.4% for tsFCI). For 20 variables tsFCI has an FDR of 88.7%, while our FDR is 2.9%. Further, the FDR for tsFCI with 100 variables is 94.1%. The ground truth for these data is 4–20 latent variables, while tsFCI finds from 33–137 latent variables, which may explain the FDR.

Scalability Our approach retains high accuracy as the number of variables increases. We further show that the approach scales well in terms of computational complexity. We use the same set-up as for the complex structures, but vary the length of the time series (2000 to 10000 by steps of 2000; and 200 to 1000 by steps of 100 to show detail). Figure 6(a) shows how run time scales linearly with the length of the time series. Even with 10000 timepoints and 100 variables, runtime is about 13min on a desktop computer with 16GB RAM. During inference of the latent time series, we maintain a table storing computed probabilities, to avoid redundant computations. Figure 6(b) shows that in practice, with >400 timepoints, we do only 1% of the theoretical maximum number of probability inferences (avoiding 99%). With 100 variables and 100 timepoints, we still avoid 80% of the max computations.

4.2. Real-world diabetes data

We now apply our approach to real-world data from individuals with T1D.



(a) Run time for reconstructing latent series

(b) Percent of computation avoided with caching

Figure 6: Evaluating scalability with (a) run time and (b) percentage of computations that are avoided due to our caching of inference results.

4.2.1. DMITRI dataset

We test our method using the Diabetes Management Integrated Technology Research Initiative (DMITRI) dataset (Feupe et al., 2013). DMITRI includes data for 17 people (10 male, 7 female) with T1D. The continuously collected data include: glucose (Dexcom 7+ CGM), insulin basal and bolus rate (insulin pump), heart rate (Polar chest strap), activity (BodyMedia SenseWear, Respironics Actiwatch), temperature (SenseWear), and sleep (Zeo Personal Sleep Coach). Since DMITRI has a high rate of missing CGM data, we use the method introduced by Rahman et al. (2015), and previously applied to this dataset, to impute missing BG values when there is a gap of less than 30min.

4.2.2. Methods

For this experiment, meal information is latent for all subjects, though activity data is measured. We discretize the data as follows. For blood glucose (BG) we set BG$<$ 70mg/dL as hypoglycemia, $70 \leq$ BG ≤ 150 as euglycemia, and BG$>$ 150 as hyperglycemia. We discretize the activity data into no exercise (resting), moderate exercise, and intense exercise using both the heart rate (HR) and the METs (metabolic equivalents) recorded by the SenseWear activity monitor. For METs, each 5-minute interval was discretized using ranges of under 3.0 (sedentary), 3.0-6.0 (moderate), 6.0-9.0 (vigorous) and above 9.0 (very vigorous). HR is discretized into three zones using age and baseline resting heart rate (HRrest). Maximum HR (HRmax) is measured as 220-age, and the three zones (zone 1, 2, and 3) of HR are defined using: $X*$(HRmax-HRrest)+HRest, where $X \in [.5, .85, 1]$. We combined heart rate and METs as follows: resting (METs sedentary and HR in zone 1), moderate exercise (METs moderate, and HR is in zone 2), and intense exercise (METs vigorous or very vigorous, and HR in zone 3). For insulin we use presence or absence of a bolus at each time. Prior knowledge K is that meal can cause hyperglycemia in 15-45min. All other settings for both our approach and tsFCI are as in the simulated data experiments.

4.2.3. Results

We find a number of causal relationships involving exercise, meals, and glycemia: moderate exercise causes a meal in 60-85min, moderate exercise causes hypoglycemia in 70-90min, and intense exercise causes hyperglycemia in 10-20min. In comparison, tsFCI infers 36 latent variables and only finds hyperglycemia causing itself in 10-15min, and hypoglycemia causing itself in 5-10min. Our finding about the relationship between intense activity and hyperglycemia overlaps the 15-30min time window for the same causal relationship identified in (Heintzman and Kleinberg, 2016), and which is supported by prior work on activity and glycemia (Riddell and Perkins, 2006). The other two relationships go beyond that inferred in prior work on this data and the prior knowledge K. They demonstrate how our approach can be used to make novel inferences about latent variables, and both findings are supported by the literature. Moderate activity can result in hypoglycemia, and the effect of activity can persist for a prolonged duration (Basu et al., 2014), making both the inferred relationship and time window likely. Further, our finding that moderate activity leads to meals but intense activity does not makes sense as a glucose regulation strategy given the different effects of each type of activity. In terms of recovering meals, our approach identifies 67 meals in total. We consider a meal plausible if 1) the meal starts before an increase of BG, and 2) BG increases by 4mg/dL within 30 minutes after the meal starts. The threshold of 4mg/dL is commonly used in works on identifying meal onset from CGM data (Lee et al., 2009; Xie and Wang, 2015). Using this threshold, our true positive rate (accepted meals out of all detected) is 0.6, though it is possible others are true meals that simply do not meet this operational criteria. Overall our results show that prior knowledge can be leveraged to discover new relationships in observational data.

5. Conclusion

While causal inference from observational data is of increasing importance, it brings new challenges in missing data and confounding due to latent variables. In this work, we propose a new approach for causal inference from time series data with latent variables. By leveraging prior knowledge, which may come from inferences from similar datasets, other time periods, or domain expertise, we are able to reconstruct the timing of latent variables and efficiently use this to avoid confounding, even when the prior knowledge has errors or is incomplete. We demonstrate that this approach can handle canonically difficult cases (hidden common causes) and more complex structures, with higher accuracy than the state of the art. In application to real-world diabetes data, we are able to infer more causal relationships the state of the art. In future work we aim to extend this approach to cases where prior knowledge is incomplete. Code is available at: https://github.com/health-ai-lab/latent-k.

Acknowledgments

This work was supported in part by the NIH under award number R01LM011826, the James S. McDonnell Foundation, and NSF under award number 1347119.

References

R. Basu, M. L. Johnson, Y. C. Kudva, and A. Basu. Exercise, Hypoglycemia, and Type 1 Diabetes. *Diabetes Technology & Therapeutics*, 16(6):331–337, 2014.

G. Borboudakis and I. Tsamardinos. Incorporating Causal Prior Knowledge as Path-Constraints in Bayesian Networks and Maximal Ancestral Graphs. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

G. Borboudakis and I. Tsamardinos. Scoring and Searching over Bayesian Networks with Causal and Associative Priors. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

X.-W. Chen, G. Anantha, and X. Lin. Improving Bayesian Network Structure Learning with Mutual Information-Based Node Ordering in the K2 Algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):628–640, 2008.

D. Colombo and M. H. Maathuis. Order-Independent Constraint-Based Causal Structure. Learning. *Journal of Machine Learning Research*, 15:3921–3962, 2014.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40 (1):294–321, 2012.

G. F. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks From Data. *Machine Learning*, 9(4):309–347, 1992.

M. Eichler. Graphical Gaussian Modelling of Multivariate Time Series with Latent Variables. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

D. Entner and P. O. Hoyer. On Causal Discovery from Time Series Data Using FCI. In *Probabilistic Graphical Models*, 2010.

S. F. Feupe, P. F. Frias, S. C. Mednick, E. A. McDevitt, and N. D. Heintzman. Nocturnal continuous glucose and sleep stage data in adults with type 1 diabetes in real-world conditions. *Journal of Diabetes Science and Technology*, 7(5):1337–1345, 2013.

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

N. Heintzman and S. Kleinberg. Using uncertain data from body-worn sensors to gain insight into type 1 diabetes. *Journal of Biomedical Informatics*, 63:259–268, 2016.

A. Hyttinen, P. O. Hoyer, F. Eberhardt, and M. Järvisalo. Discovering Cyclic Causal Models with Latent Variables: A General SAT-Based Procedure. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

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

H. Lee, B. A. Buckingham, D. M. Wilson, and B. W. Bequette. A closed-loop artificial pancreas using model predictive control and a sliding meal size estimator. *Journal of Diabetes Science and Technology*, 3(5):1082–1090, 2009.

K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkley, 2002.

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

V. K. Raghu, J. D. Ramsey, A. Morris, D. V. Manatakis, P. Sprites, P. K. Chrysanthis, C. Glymour, and P. V. Benos. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *International Journal of Data Science and Analytics*, 6(1):33–45, 2018.

S. A. Rahman, Y. Huang, J. Claassen, N. Heintzman, and S. Kleinberg. Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. *Journal of Biomedical Informatics*, 58:198–207, 2015.

M. C. Riddell and B. A. Perkins. Type 1 Diabetes and Vigorous Exercise: Applications of Exercise Physiology to Patient Management. *Canadian Journal of Diabetes*, 30(1):63–71, 2006.

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

L. Song, M. Kolar, and E. P. Xing. Time-Varying Dynamic Bayesian Networks. In *NIPS*. 2009.

P. Spirtes. An Anytime Algorithm for Causal Inference. In *AISTATS*, 2001.

Peter Spirtes, Clarke Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, second edition, 2000.

M. Voortman, D. Dash, and M. J. Druzdzel. Learning Why Things Change: The Difference-Based Causality Learner. In *UAI*, 2010.

J. Xie and Q. Wang. Meal Detection and Meal Size Estimation for Type 1 Diabetes Treatment: A Variable State Dimension Approach. In *ASME Dynamic Systems and Control Conference*, 2015.

J. Zhang. On the Completeness of Orientation Rules for Causal Discovery in the Presence of Latent Confounders and Selection Bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

N. L. Zhang. Hierarchical Latent Class Models for Cluster Analysis. *Journal of Machine Learning Research*, (5):697–723, 2004.