Causal Discovery in Financial Markets: A Framework for Nonstationary Time-Series Data

Agathe Sadeghi*
Stevens Institute of Technology
Hoboken, NJ, USA
asadeghi@stevens.edu

Achintya Gopal
Bloomberg
New York, NY, USA
agopal6@bloomberg.net

Mohammad Fesanghary
Bloomberg
New York, NY, USA
mfesanghary1@bloomberg.net

Thursday 28^{th} December, 2023

Abstract

A deeper comprehension of financial markets necessitates understanding not only the statistical dependencies among various entities but also the causal dependencies. This paper extends the Constraint-based Causal Discovery from Heterogeneous Data algorithm to account for lagged relationships in time-series data (an algorithm we call CD-NOTS), shedding light on the complex causal relations between different financial assets and variables. We compare the performance of different algorithmic choices, such as the choice of conditional independence test, to give general advice on the effective way to use CD-NOTS. Using the results from the simulated data, we apply CD-NOTS to a broad range of indices and factors in order to identify causal connections among the entities, thereby showing how causal discovery can serve as a valuable tool for factor-based investing, portfolio diversification, and comprehension of market dynamics. Further, we show our algorithm is a more effective alternative to other causal discovery algorithms since the assumptions of our algorithm are more realistic in terms of financial data, a conclusion we find is statistically significant.

1 Introduction

Actions have consequences; that explains causality in simple words. Going beyond correlation or association, it is a guide which helps us gain a deeper understanding of the relationships between events, allowing us to perceive the impact of one phenomenon on another. An empirical causal framework would consist of three phases: (a) causal discovery; (b) causal inference; (c) causal explainability. In the first phase, the causal network is estimated from the data based on independence tests. When the network is discovered, then different scenarios can be tested on the structure. Afterwards, we can dive deep into the intuitive why of the causal network and results.

In this paper, our focus is on causal discovery, which aims to uncover the causal dependency structure from the observed data (Spirtes et al., 2001). While the traits in the data narrow down the possible causal network structures, it may not necessarily be sufficient to discover the complete underlying true causal network.

Most causal discovery methods can be categorized into three types of algorithms: constraint-based (e.g., Spirtes et al. (2000); Huang et al. (2020)), score-based (e.g., Chickering (2003); Silander and

*work performed while interning at Bloomberg.

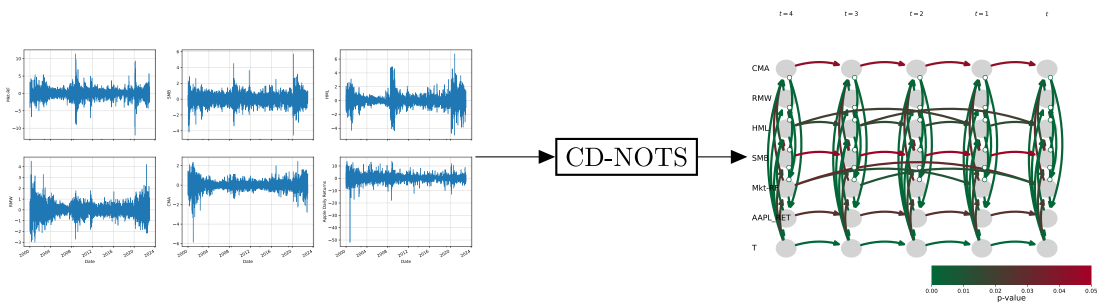


Figure 1: We show a high-level diagram of the problem we are solving. We insert some time-series data into our causal discovery algorithm, from which we get a causal diagram of how the variables are related.

Myllymäki (2012); Lam et al. (2022)), and functional causal model-based (FCM) (e.g., Shimizu et al. (2006, 2011); Hyvärinen et al. (2010)). The constraint-based causal discovery algorithm infers causal dependency between nodes by running statistical independence tests among conditional and marginal probabilities to establish the network structure. The score-based algorithm defines a score criterion (e.g., likelihood scores Schwarz (1978); Buntine (1991)) and optimizes for the causal structure that best fits the data. On the other hand, FCMs intend to model the functional mechanisms of the variables. This calls for additional assumptions on the distribution of the data or domain-specific knowledge.

Causal discovery algorithms require assumptions about the data generating process; thus, each algorithm has limitations depending on the realism of the assumptions. For example, some only work on linear systems (Granger, 1969), some assume a Gaussian distribution for the variables (Ramsey and Andrews, 2017), and many assume there are no latent (unobserved) confounders (Zhang and Hyvärinen, 2009). Further, implementing causal discovery algorithms on time-series data adds additional challenges, such as sampling interval and nonstationarity, to the existing issues. There are ways to reduce the nonstationarity of time-series data such as using stock returns as opposed to price. However, empirically, we find this is insufficient to make the process stationary. Unlike evaluation of predictive performance, in general, causal discovery algorithms, when applied in practice, do not have ground truth data or ways to test the accuracy of the discovered network via randomized controlled trials. This leads to the importance of ensuring that the assumptions of the causal discovery algorithm pertains to the situations in which they are used. In our case studies (Section 5), we show that, in the financial examples we experiment with, the data is neither linear, Gaussian, nor stationary.

Our developed algorithm, Constraint-based Causal Discovery from Heterogeneous Time-Series (CD-NOTS), overcomes the limitations mentioned above. It is a nonparametric approach, meaning it can capture both linear and non-linear causal relations and can also capture non-Gaussian distributions. It is capable of handling nonstationarity, and it is able to detect both lagged and contemporaneous causalities. Prior work has developed and used causal discovery algorithms that capture subsets of these properties; our work is the first (to the best of our knowledge) to tackle all three. We then analyze the patterns found in financial data using this algorithm.

The structure of this paper is as follows. In Section 2, we outline the related work and literature. Moving on to Section 3, we explain our methodology in detail. In Section 4, we conduct an investigation of simulated and real financial data to assess the framework's validity and performance. Finally, in Section 5, we summarize the key findings of the paper and propose potential avenues for future research.

The most similar to our work is Huang et al. (2020), which we extended to handle lagged relationships in time-series data. In Section 3.3, we show a simple example in which we illustrate the difference in the causal relationships discovered when accounting for lagged relationships. Huang et al. (2020) discusses in their work a way in which their algorithm can be extended to lagged relationships; however, in Section 3.4, we show a simple example in which their extension would discover an incorrect graph.

2 Related Work

Causality explores the mechanisms that influence the outcomes, providing a more comprehensive picture of the why and how of unfolding events. The Rubin Causal Model (Rubin, 1974) and Pearl's Structural Causal Model (SCM) (Pearl, 2009a) are two approaches that are mainly considered when talking about causal relations.. A Structural Causal Model (Pearl, 2014) defines the causal relationships between variables via functions, taking the cause and the effect variables as input and output respectively. The SCM can be represented as a graphical model which can then be used for causal inference. On the other hand, the Rubin model (Splawa-Neyman et al., 1990), also known as the Potential Outcomes Framework, defines the conditional independence properties of the treatment assignment mechanism in order to analyze the unit-based response variable. Often, this framework requires a set of covariates that satisfy conditional ignorability, or in other words, no unobserved confounders. Note that prior work has shown that SCMs are formally equivalent to the Potential Outcomes Framework (Pearl, 2009b).

Many statisticians have formulated multiple computational techniques and procedures to discover causal links through learning the causal relationships, typically depicted as directed acyclic graphs, as reviewed in Glymour et al. (2019). After obtaining the causal skeleton, there is an attempt to orient the edges. The two main methods are constraint-based and score-based. Constraint-based ones can handle confounders as the score-based approach is applicable only when there is a no-confounder assumption.

Constraint-based algorithms run independence-based tests to establish the underlying causal network (Nogueira et al., 2022). This set of algorithms starts from a full graph, orients edges based on conditional independence (CI), and identifies colliders. It is an iterative process which ends when there are no further edges to orient. Two widely used algorithms are Peters and Clark (PC), and Fast Causal Inference (FCI) (Spirtes et al., 2000). PC assumes the absence of confounders, which results in asymptotically correct causal information that helps with simplifying the problem. There have been improvements to the PC algorithm that tend to reduce the number of conditional independence tests and relax the assumptions (e.g., Tsamardinos et al. (2006); Bühlmann et al. (2009); Ramsey et al. (2012)). On the other hand, FCI produces asymptotically correct outcomes even when confounders are present; Colombo et al. (2012) and Rohekar et al. (2021) are two improved FCI works. Since our work is an extension of the PC algorithm, we utilize four different tests: Kernel-based Conditional Independence Test (KCIT) (Zhang et al., 2012), Randomized Conditional Correlation Test (RCoT) (Strobl et al., 2017), Conditional Mutual Information K-nearest neighbor (CMIknn) Test (Runge, 2017), and Partial Correlation (ParCorr) (Baba et al., 2004).

More recently, causal structure discovery algorithms have been developed that do not require statistical tests but instead use assumptions of the functional form or of the distribution in order to recover the structure (Shimizu et al., 2006; Zhang and Hyvärinen, 2010; Zhang and Hyvarinen, 2012). For example, these algorithms assume absence of confounders and a non-Gaussian distribution within the data, enabling them to uncover the asymmetry inherent in the causes and effects distribution (Gendron et al., 2023).

Causal discovery in time-series data is about uncovering and quantifying causal relationships among variables that evolve over time. The difference between causal discovery methods and traditional time-series analysis is that causal discovery aims to identify the cause and effect relationships in the data, while traditional time-series analysis is more focused on prediction.

Granger causality (Granger, 2001) is an approach that estimates the linear causal relations using a VAR model (Hyvärinen et al., 2010). There is additional work for causal structure recovery using Dynamic Bayesian Networks and Hidden Markov models (Moraffah et al., 2021). Another class of causal discovery methods uses information theoretic quantities such as Mutual Information (Palus et al., 2001), and Transfer Entropy (Schreiber, 2000). Furthermore, graphical approaches like Time-Series Fast Causal Inference (Entner and Hoyer, 2010) and PCMCI (Runge et al., 2019) are more recent graphical approaches that are tweaked to be able to handle time-series data.

Most of the time-series causal discovery literature mentioned are focused on stationary time-series, and their approaches might lead to spurious causal relations if the time-series contains distributional shifts. Some work has been developed to address this gap (e.g., Peters et al. (2015); Zhang et al. (2017)).

One approach is to assume the consistency of the nonstationarity reflected in the network across time. For instance, Huang et al. (2020) proposes an algorithm for heterogeneous time-series that only checks the nonstationarity at one time t. Our developed algorithm extends this to handle nonstationarity in lagged relationships.

3 Methodology

Time-series is the fundamental type of data used in many domains, from finance to climate science and engineering, requiring customizing algorithms to take this type of data as input. This tailoring requires taking into account some of the characteristics that this type of data carries, such as temporality, lead-lag dependency, and nonstationarity. Similar to CD-NOD, our algorithm, Constraint-based Causal Discovery from Heterogeneous Time-Series (CD-NOTS), has four steps: adding the time-indexed node, using conditional independence tests to discover the causal skeleton (undirected causal graph), orienting the edges via prior knowledge and specific structures (e.g., the arrow of time and V-structures), and, finally, orienting remaining edges based on causal change independence. By design, our framework is able to handle time-series as input and, accordingly, takes into account its unique characteristics; in comparison, CD-NOD does not handle lagged relationships (Section 3.4).

With suitable assumptions, the causal network, using only observational data, can be estimated using conditional independence tests (Schölkopf et al., 2021). We follow the same set of assumptions commonly used in the causal network field.

Assumption 1 (Causal Sufficiency)**.** We assume that any potential confounders can be expressed as a smooth function of time. Hence, the time index node is a representation of all confounders.

Assumption 2 (Causal Faithfulness)**.** We assume that the conditional independence between the observed data is a true reflection of the absence of a direct relation in the variables given the conditional set. In other words, if X is conditionally independent of Y given Z ($X \perp\!\!\!\perp Y|Z$), then this is true in the underlying causal network.

Assumption 3 (Randomness)**.** The data points are randomly selected from the population implied by the causal model (i.e., there is no selection bias).

Assumption 4 (Causal Consistency)**.** We assume that the causal relations between the variables, including the time indexed node, are consistent through time. This means that when lagged relations are present, the contemporaneous causal relations are the same in each lag and the cross lag causal relations also repeat themselves.

3.1 CD-NOTS Algorithm

We extend the algorithm described in Huang et al. (2020) in order to identify the causal network for time-series. Our developed approach, termed Constraint-based Causal Discovery from Nonstationary Time-Series (CD-NOTS), enhances the original method for identifying causal relationships between time-series. We show a high-level diagram of our algorithm in Figure 2.

Stage 1 Assume we have N time-series. Within our network setup, each time-series is treated as a distinct node, denoted as $V_{i,t}$, representing the time-series i at time t. To handle nonstationarity, we introduce a time-indexed node, U_t. Specifically, for any nonstationary time-series i, we anticipate a connection between $V_{i,t}$ and U_t across all time points. We start from a full graph including the time-indexed node, where all the edges are drawn out.

Stage 2 With Assumptions 1, 2 and 3 in place, we reconstruct the causal skeleton. Employing CI tests, we investigate the independence between a node and the time-indexed node given a subset of nodes S. This process figures out the edges between nodes U_t and $V_{i,t}$. The hypothesis test is:

$$H_0 : P(V_{i,t}, U_t|S) = P(V_{i,t}|S)\, P(U_t|S)\,,$$
$$H_1 : P(V_{i,t}, U_t|S) \neq P(V_{i,t}|S)\, P(U_t|S)\,,$$

which we test for all subsets $S \subseteq \{V_{k,t-l} \mid k = 1, ..., N; l = 0, ..., L\} \setminus \{V_{i,t}\}$ where L denotes the maximum considered lag. Note that this hypothesis examines the relationship between the time-indexed node and other nodes at the identical time point t, maintaining consistency across time $t - l$.

Afterwards, utilizing the same test, we evaluate the independence of node pairs $V_{i,t-l}$ and $V_{j,t}$ excluding U_t. The null hypothesis is that the pair are conditionally independent, implying that their conditional probability equals the product of their marginals. If the obtained p-value falls below the confidence level, signifying dependence, the edge between the pair remains. The hypothesis test is:

$$H_0 : P(V_{i,t-l}, V_{j,t}|S) = P(V_{i,t-l}|S)\, P(V_{j,t}|S)\,,$$
$$H_1 : P(V_{i,t-l}, V_{j,t}|S) \neq P(V_{i,t-l}|S)\, P(V_{j,t}|S)\,,$$

where $S \subseteq \{U_t\} \cup \{V_{k,t-m} \mid k = 1, ..., N; m = 0, ..., L\} \setminus \{V_{i,t-l}, V_{j,t}\}$. It is worth noting that when $l = 0$, indicating no lag, we capture the contemporaneous relationships between nodes. On the other hand, when $l > 0$, we discover the lagged dependencies among the time-series. Under Assumption 4, we maintain the structure consistent over time. This means that the discovered dependencies between $V_{i,t-l}, V_{j,t}$ remain the same for $V_{i,t-l-m}, V_{j,t-m}; m = 1, ..., l - 1$. The discovered network is undirected and has the edge minimality condition. This condition signifies that the network has the smallest number of causal changing mechanisms. This is due to the faithfulness across the entire network (Ghassami et al., 2018).

Stage 3 Once we have derived the network skeleton, our next step involves orienting the edges based on existing knowledge. If there are any nonstationary nodes, meaning there exists an edge between the node and time indexed node, then the orientation is from the time indexed node to the nonstationary node, as time is what causes the node's corresponding time-series to be nonstationary. Moreover, all the edges where $l < m$ between pairs $V_{i,t-l}$ and $V_{j,t-m}$, the orientation is from $V_{i,t-m}$ to $V_{j,t-l}$, as logically the effect would happen after the cause. These directed edges along with the conditional independence tests can be used to orient other edges using the unique properties of V-structures (colliders) (Spirtes et al., 2000; Meek, 1995).

Stage 4 There might still remain some edges that are not yet directed. In the fourth stage, we rely on the independent changes of the causal modules (Pearl, 2000). A dependency measure is defined as an extension of the Hilbert Schmidt Independence Criterion (Gretton et al., 2008), which shows the level of dependency between causal modules. This is done by developing a kernel embedding of nonstationary conditional distributions and using their Gram matrix to create a test statistic (Huang et al., 2017). We apply the Meek orientation rule (Meek, 1995) on top of this to orient other edges. Note that there may remain edges that are undirected even after the third and fourth stage; in this scenario, no comments can be given on the causal direction of the corresponding pair of nodes.

3.2 Testing the Assumptions

The main assumptions we claim that are incorrectly assumed in prior causal discovery algorithms is stationarity, no lagged relationships, and linearity. Given the fact our algorithm does not make the first two assumptions, we can analyze the graph from our algorithm to see if nonstationary and lagged relationships are found; if found, then the conclusion is that, assuming stationarity (lagged relationships) is the null hypothesis, the statistical tests could not reject the null hypothesis.

On the other hand, for the linearity assumption, we modify an algorithm used in prior work (Peters et al., 2014) used in order to test post-hoc if the relationships are linear. Specifically, the relationships are assumed to be linear in conjunction assuming with additive noise. The algorithm used in Peters et al. (2014) is to perform a regression and then an independence test on the noise. Say the null hypothesis we are testing is that X and Y have a linear relationship given Z (where we allow for Z to have a non-linear relationship with Y and the noise to be heteroscedastic with respect to Z), we perform a Gaussian process regression $\hat{f}$, where we use the linear kernel for X and a non-linear kernel (e.g., RBF) for Z. Having done that, we then perform a CI test on $Y - \hat{f}(X, Y) \perp\!\!\!\perp X|Z$. Note that our test is more generic than testing if *all* relationships are linear.

Stage 0: Consider four time-series
A, B, C and D with their lag one

Stage 1: Causal Skeleton Initialization
Add the time-indexed node U

Stage 2: Causal Skeleton Discovery
by independence tests

Consistency of skeleton check

Stage 3: Causal Direction Identification
by prior knowledge

Consistency of directions check

Stage 4: Causal Direction Identification
by causal independent changes

Consistency of directions check

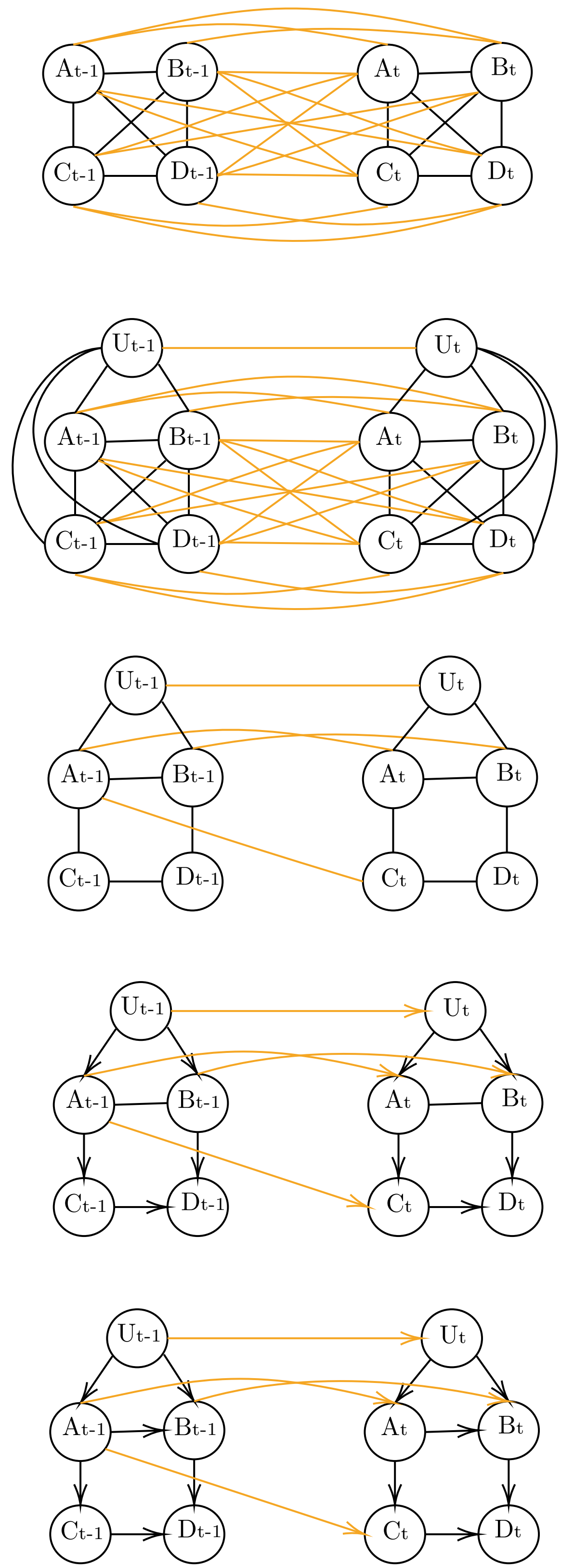


Figure 2: CD-NOTS algorithm schematic example

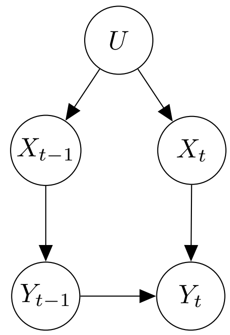


Figure 3: Diagram of example causal graph for which CD-NOD's time-series algorithm would make a mistake.

3.3 Importance of Modeling Lagged Relationships

We show with a toy example how, if one were to not model lagged relationships but the data contains them, the graph would find nonstationary variables. Say we have one variable X_t where:

$$dX_t = -x_t dt + \sqrt{2} dW_t \,,$$
$$X_0 = 0 \,.$$

To discretize, we set t to be integers. Since this is simply an Ornstein-Uhlenbeck (OU) process, we have:

$$P(x_t | x_u) = P(x_t | \mathcal{N}(x_u e^{-|t-u|}, e^{-|t-u|})) \,,\; u < t \,.$$

From this, we can see that $P(x_t|x_0)$ is a time dependent distribution whereas $P(x_t|x_{t-1})$ is not. This implies if we ran CD-NOTS assuming no lagged relationships (i.e., CD-NOD), x_t would be considered nonstationary; whereas, if we ran CD-NOTS assuming a single lag, then $X_{t-1} \to X_t$ and X_t would not be considered nonstationary.

3.4 Comparison to Algorithm in CD-NOD

Huang et al. (2020) presents an algorithm to extend CD-NOD to time-series. However, we find that in the first step in the algorithm ('Detection of changing modules') to remove connections from the time varying node U and observed variables $V_{i,t}$), the tests are only using a subset of $\{V_{j,t}\}_{j=1}^{N} \setminus U$. We can see in Figure 3 that $Y_t \not\perp\!\!\!\perp U$ and $Y_t \not\perp\!\!\!\perp U | X_t$. This implies that the algorithm presented in Huang et al. (2020) would leave out the edge from U and Y_t.

4 Evaluation

In our algorithm, one major design decision is the conditional independence test. In order to ascertain which conditional independence test might be most effective, we test our algorithm on different simulated datasets. In total, we tested graphs with 3, 4, 5, 6, 8, 10, and 15 nodes, for a total of 350 graphs (50 graphs per number of nodes). For each graph, we tested having 50, 150, 300, 500, and 1,000 data points.

For the conditional independence tests, we specifically test using partial correlations (ParCorr, a conditional independence test that assumes linearity), Kernel-based Conditional Independence Test (KCIT), Randomized Conditional Correlation Test (RCoT), and Conditional Mutual Information K-nearest neighbor (CMIknn) Test. For both KCIT and RCoT, there are a few variations in how the p-value can be computed, specifically using the Satterthwaite–Welch (SW) method (Welch, 1938; Satterthwaite, 1946) or Hall– Buckley–Eagleson (HBE) method (Hall, 1983; Buckley and Eagleson, 1988). The goal of these algorithms is to approximate the CDF of a weighted sum of squared normals, which is the limiting behavior of the test statistic for KCIT and RCoT. The hyperparameters used for each test can be found in Appendix A.

In Figure 4, we compare the average F-score for each CI test across a varying number of nodes in the graph and a varying number of data points per graph. As expected, the F-score improves as the number of data points increase. Possibly most interesting is that ParCorr performs the best for low data regimes. Further, we can see, as found by previous work (Bodenham and Adams, 2015), that the SW variant of KCIT tends to be weaker than HBE (except when there are a large number of nodes and only 50 data points). However, there seems to be little difference for RCoT. So, relying on results in Bodenham and Adams (2015), we recommend the HBE variant. Finally, while CMIknn is quite competitive, we do not recommend its usage due to its lengthy runtime (Figure 5).

Comparing against the best tests (KCIT HBE, RCoT HBE, ParCorr, CMIknn), we consider PCMCI (Runge et al., 2019) as a benchmark since it is a comparable PC-based algorithm and is widely used

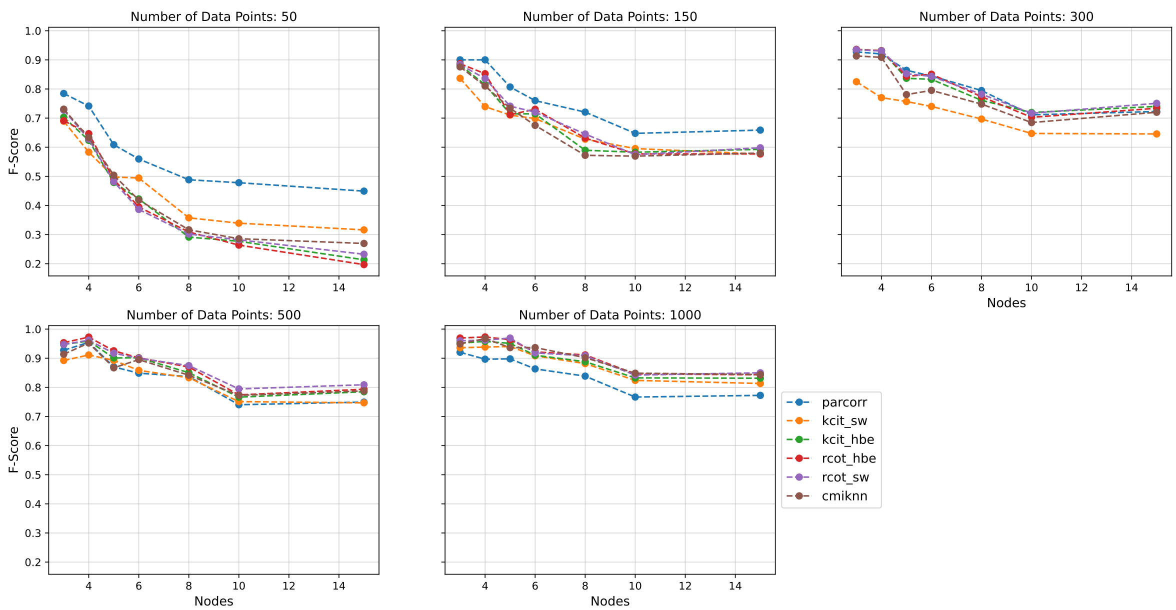


Figure 4: F-score evaluation for CD-NOTS with different CI tests, tested out on many different simulated datasets.

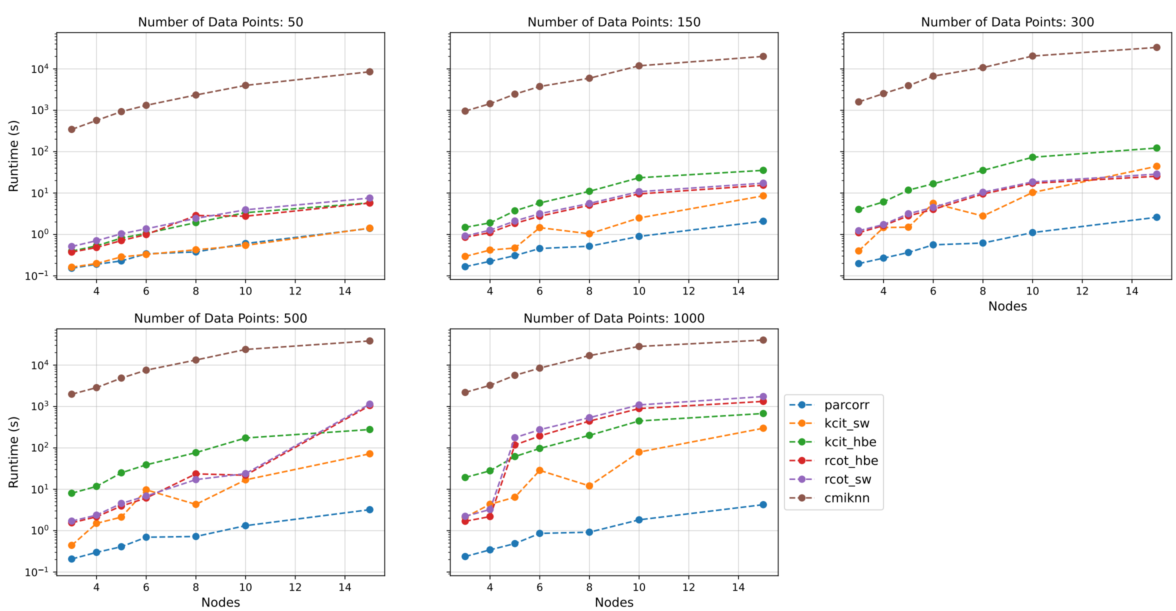


Figure 5: Runtime evaluation for CD-NOTS for different CI tests. Note the y-axis is in log scale.

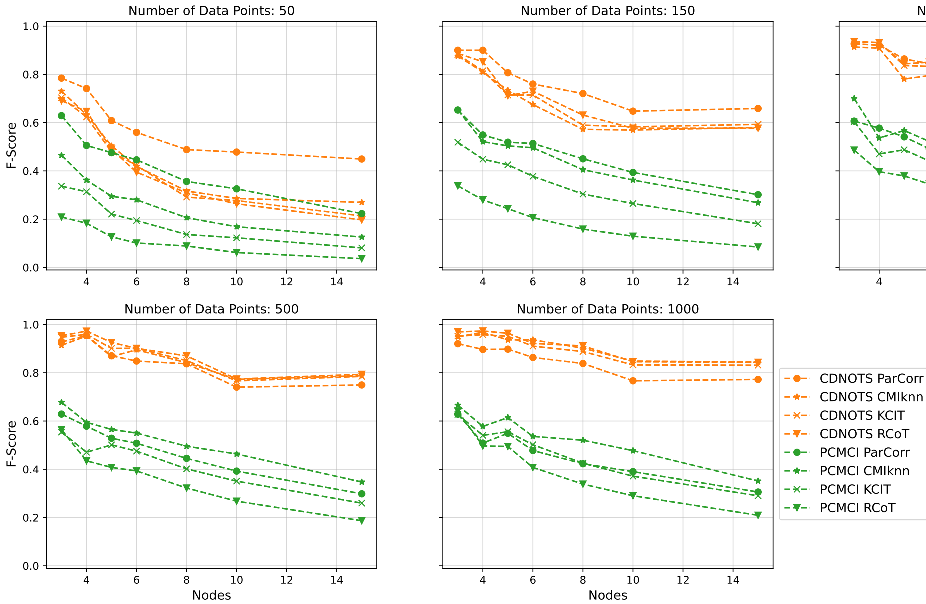


Figure 6: F-score evaluation, comparing CD-NOTS and PCMCI, tested out on many different simulated datasets.

in the causal discovery literature. The main difference in PCMCI over prior PC algorithms is its pruning of the set of variables which to condition on in the CI test. In Figure 6, we see that CD-NOTS consistently outperforms PCMCI.

5 Case Studies

We apply CD-NOTS on three different datasets: 1) Fama-French factors and Apple's returns, 2) unemployment, CPI, and PPI of different countries, and 3) Price-to-Book ratio and stock returns of financial companies in the S&P 500. Most of the work done in factor investing and economic indices' analysis rely on the correlation between the variables and do not consider the fundamental question of "why" (López de Prado, 2023). In each case study, we explain the data and properties we see in the data or has been found in prior work, discuss the implications of the causal networks found by CD-NOTS, and, finally, show that the assumptions of stationarity, linearity, and no lagged relationships would be incorrect to assume and thus, any algorithm making said assumptions would be inaccurate and invalid to use on these datasets.

For all the graphs in the section, we color by the maximum p-value for that edge across all tests run during the causal skeleton step (Stage 2). While we recognize that p-values are not a metric to measure the strength of the relationship, we can think of the vicinity of the p-value to 0.05 as a measure of the sensitivity of the graph to that hyperparameter (i.e., the threshold).

5.1 Fama-French Factors

Data For our first case study, we analyzed the relationship between the Fama-French factors (Fama and French, 1993) and Apple's returns from the beginning of 2000 to the end of 2022. In the data in Figure 7, we show that there is some correlation between the variables (e.g., the spike in volatility around 2008 can be seen in all six subplots) and, accordingly, all variables show volatility clustering.

Results and Discussion In Figure 8, we ran CD-NOTS using KCIT (RCoT results are in Appendix B) on three non-overlapping periods of six to seven years (approximately 2000 observations each). We note in all three periods that there is no connection from T to Apple's returns (AAPL_RET), implying that the nonstationarity observed in Apple's returns can be explained away by the Fama-French factors. Further, the lagged relationships tends to be sparse, with many of the found lagged

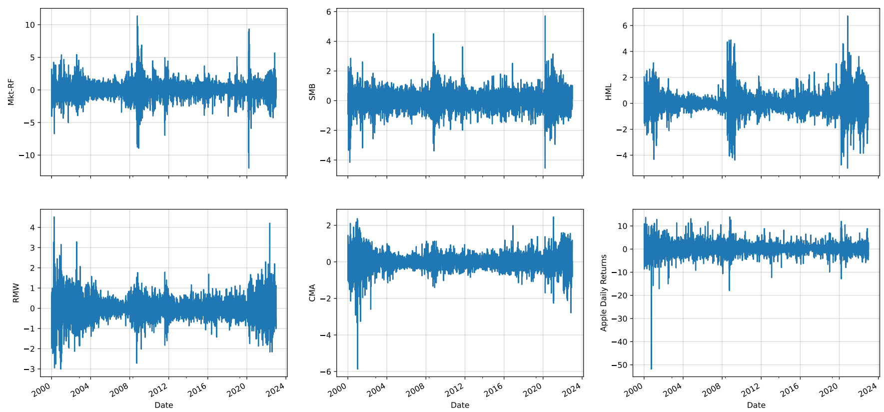
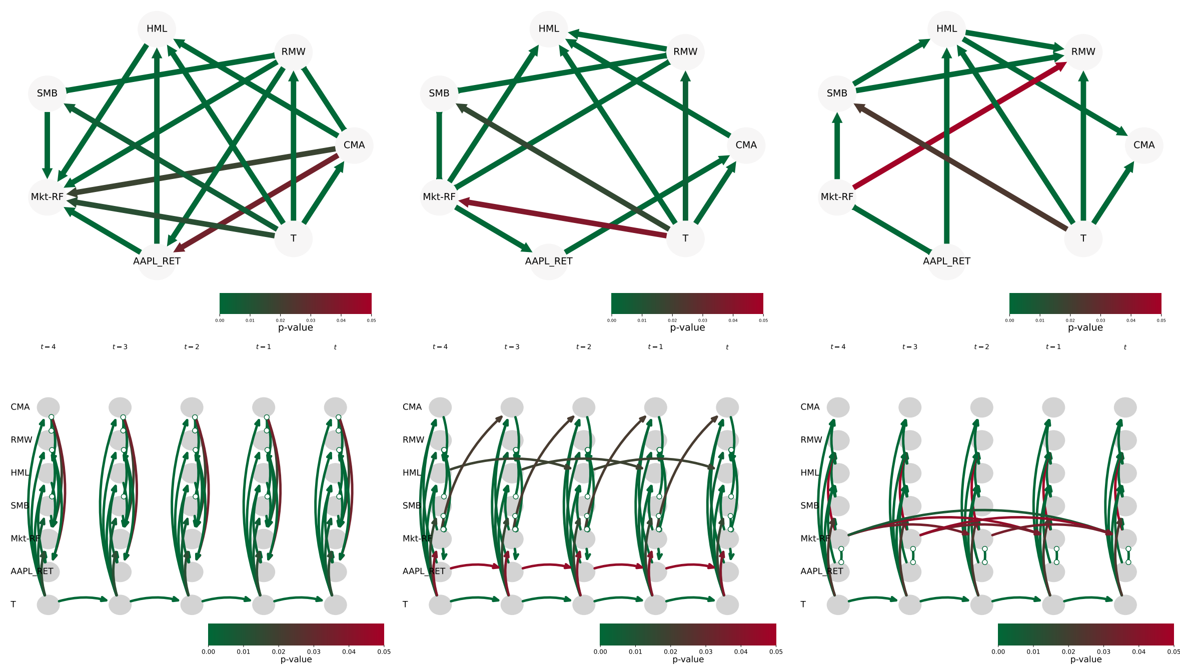


Figure 7: Time-series of Fama-French factors and Apple's returns.



(a) Jan 2000 - Dec 2007 (b) Jan 2008 - Dec 2015 (c) Jan 2016 - Dec 2022

Figure 8: Results of CD-NOTS KCIT run on three non-overlapping time periods.

relationships having a high p-value (close to 0.05). However, we see that the relationship between the factors seems to change over time, as well as the relationship between the factors and Apple's returns. In Figure 9, we run CD-NOTS with KCIT on the whole time range, and we can see now that that there is a connection from T to Apple's returns (AAPL_RET). This is to be expected for two reasons: 1) over the course of the last 22 years, Apple's business model has changed, leading to exposures to different factors over time, and 2) the changing relationship between the factors and Apple's returns, as seen in Figure 8.

Evaluating Assumptions One of the assumptions of CD-NOTS is causal consistency that the causal graph does not change over time. We assert that the change in dependencies between the factors and Apple's returns that we see in Figure 8 are not a result of changes in the causal directions but in the strength of the causal relationships as captured by the nonstationarity. Further, we note that the fact that there is a relationship between T and all the Fama-French factors (nonstationarity) implies that any causal discovery algorithm assuming stationarity would give invalid results. Similarly,

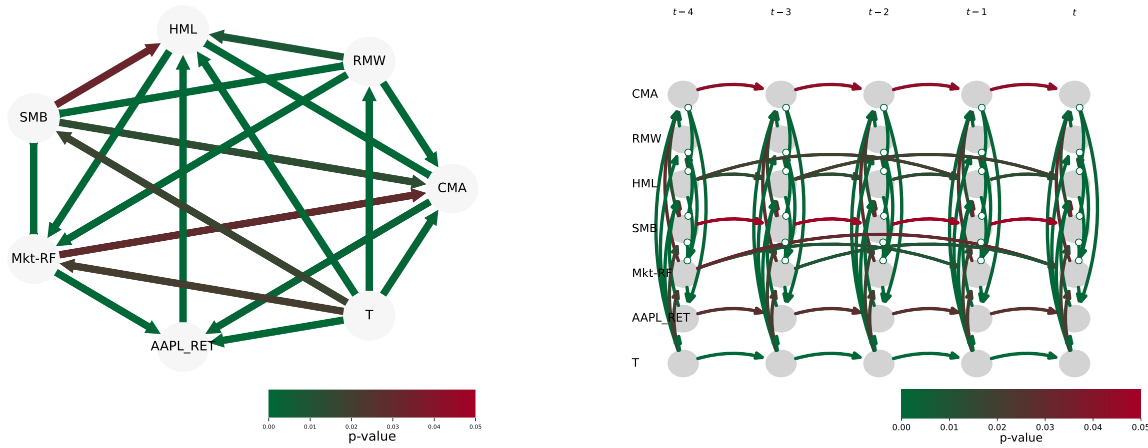
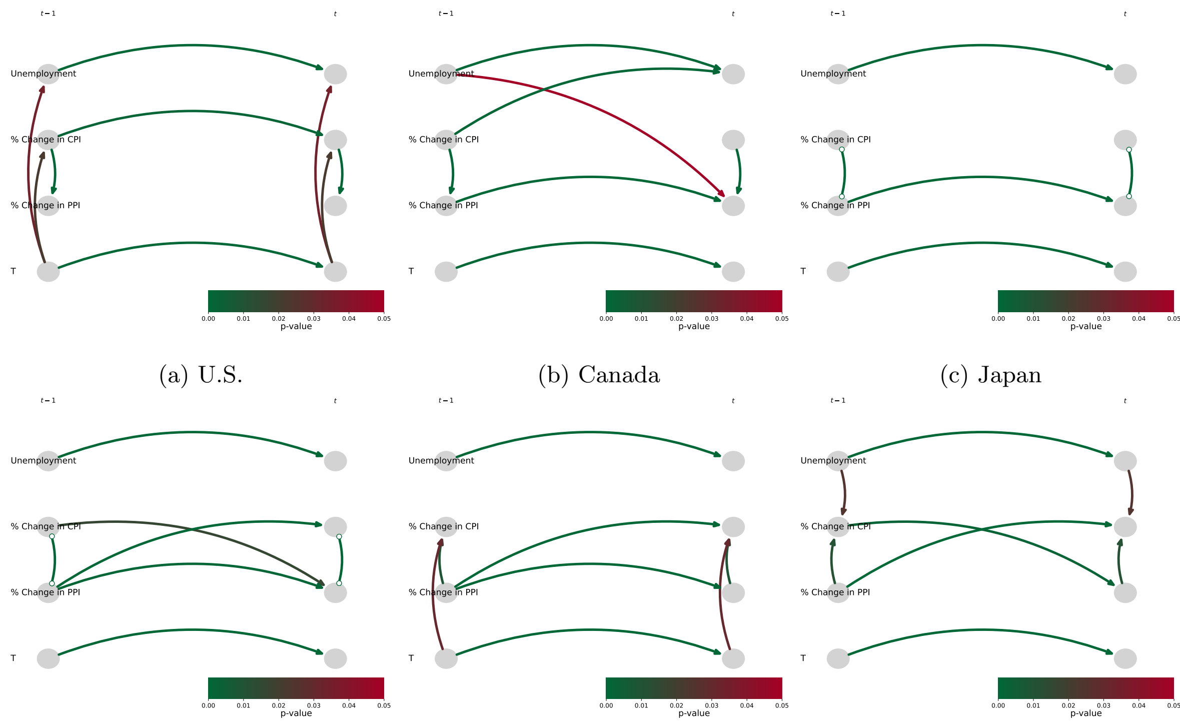


Figure 9: Results of CD-NOTS KCIT run from beginning of 2000 to end of 2022.



(a) U.S. (b) Canada (c) Japan

(d) France (e) United Kingdom (f) Italy

Figure 10: Results of CD-NOTS ParCorr run on six countries' economic variables.

we find statistically significant lagged relationships. Further, for our linearity test, we test the linearity of Apple's returns and each parent factor (CMA and Mkt-RF) and find the linearity hypothesis can be rejected for both factors; in other words, assuming linearity in the relationships between variables would be incorrect and thus could lead to incorrectly identified causal relationships.

5.2 Economic Data

Data For our second case study, we analyzed the relationship between the month-over-month percentage change in CPI, the month-over-month percentage change in PPI, and unemployment for the U.S., Japan, Canada, India, Italy, the U.K., and France. We use monthly data from the beginning of 2000 to the end of 2023.

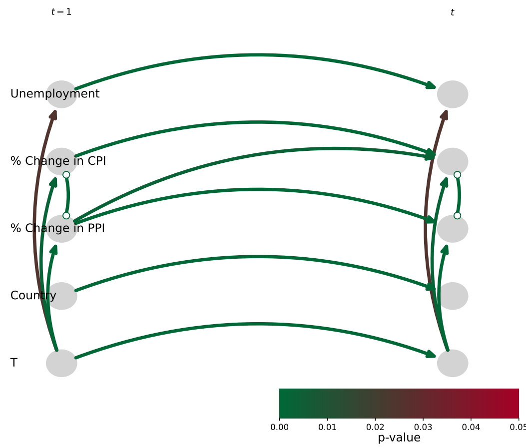


Figure 11: Results of CD-NOTS KCIT run on the economic variables of multiple countries.

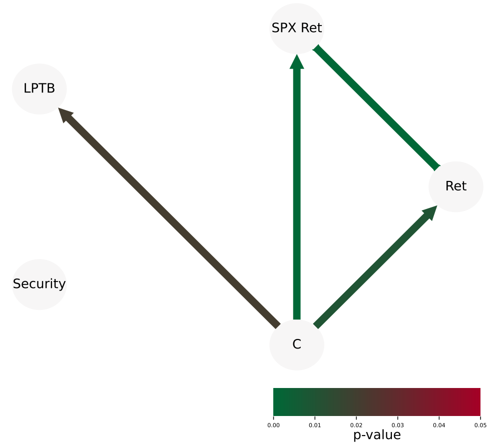
Results and Discussion In Figure 10, we run CD-NOTS with ParCorr on the U.S., Japan, Canada, Italy, France, and the U.K.; we use ParCorr since we have no more than 300 observations per country. We can see that, while the graphs are not identical, they are quite similar, especially if we ignore edges with a higher p-value. Further, we notice the three graphs for France, the U.K., and Italy are very similar, possibly due to their geographic and economic proximity. We repeat the experiment with all countries using KCIT (Figure 11); further, we introduce another static variable (Country) into the graph. Similar to the previous case study in which the graph over all time periods is akin to a superset of the edges of individual time periods, we see a similar behavior here. Note that Country does not point to any other variable suggesting that the behavior across countries is independent of the country; this observation could be used to justify training a single econometric model across countries.

Evaluation Assumptions Similar to the previous case study, we see statistically significant lagged relationships and statistically significant nonstationarity, implying we would not be able to trust the results of an algorithm that cannot handle these types of properties. For our linearity test, we test the linearity of the percentage change of CPI and each parent node (both including and excluding the contemporaneous percentage change in PPI since its direction was not inferred). We find that the only time the linearity hypothesis cannot be rejected is when using the contemporaneous percentage change in PPI as a parent and checking the linearity with the one-lag percentage change of CPI. We note that this is inconsistent with our usage of ParCorr in Figure 10 since ParCorr is a linear-based CI test. In Appendix C, we show the results of using KCIT and RCoT where we find qualitatively that the graphs are similar.

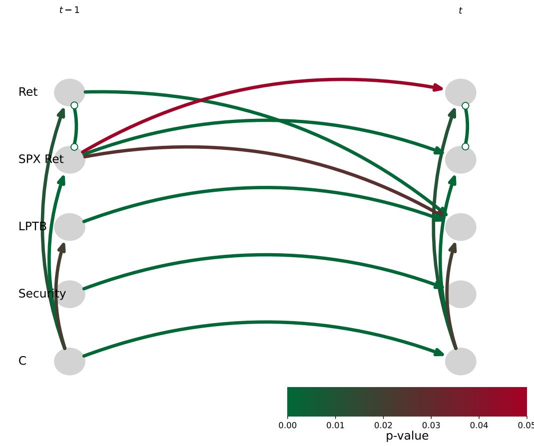
5.3 Company Financials and Returns

Data For our third case study, we analyzed the relationship between the Price-to-Book ratio and returns of financial companies in the S&P 500. We use data from the beginning of 2010 to the end of 2023. We utilize reported quarterly numbers, where we set the date to be the end of each quarter. To ensure no look-ahead bias when defining the Price-to-Book ratio, for any given date (i.e., end of quarter), we use the Price-to-Book ratio available that day. We join the next quarter returns with this company financials information. For our experiments, we use the log Price-to-Book ratio (LPTB) as well as the normalized LPTB, which we define as the LPTB normalized by the mean and standard deviation of the LPTB of financial companies in that quarter.

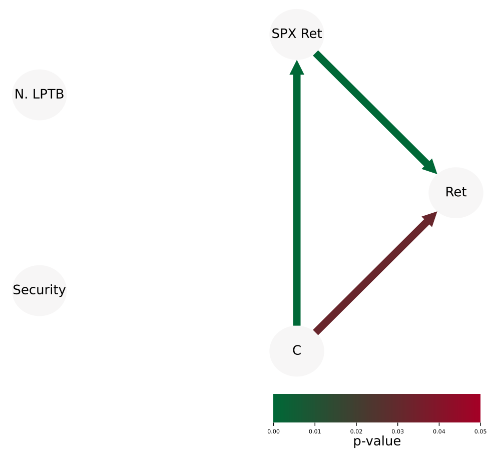
Results and Discussion In Figure 12b and Figure 12d, we compare the graphs we get running CD-NOTS using KCIT using normalized LPTB instead of LPTB. We find that the nonstationarity is removed (T does not point to N. LPTB). Further, we see that the previous quarter's normalized LPTB point to the next quarter's returns with a reasonable p-value suggesting that its inclusion in the graph is not purely a function of the threshold we chose. In Figure 12e, we compare the next quarter returns and the normalized LPTB and see that the there is a small trend in the mean (the black circles). To



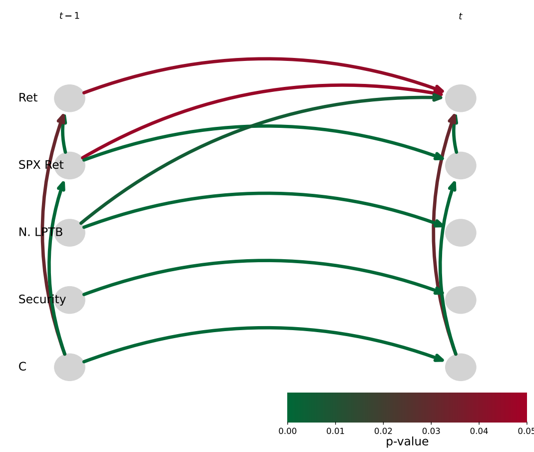
(a) Contemporaneous Effects



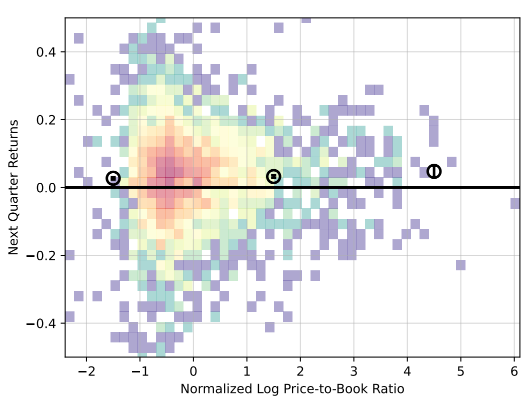
(b) Lagged Effects



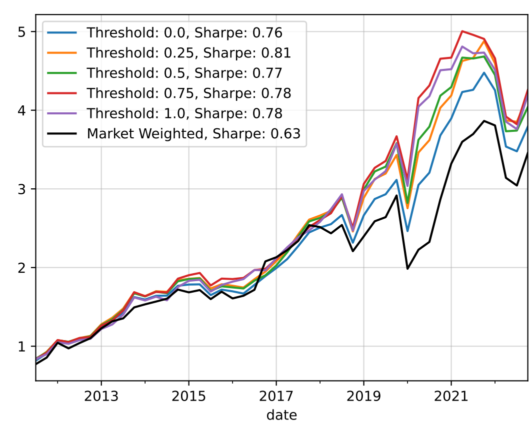
(c) Contemporaneous Effects



(d) Lagged Effects



(e) Comparison of Returns and N. LPTB



(f) Cumulative Returns of Strategies

Figure 12: Results of CD-NOTS KCIT run on financials data for log Price-to-Book ratio (LPTB) and stock returns (Ret). Normalized LPTB (N. LPTB) refers to normalizing the LPTB by the mean and standard deviation of the LPTB of financial companies in that quarter. For Figure 12f, our strategy is long-ing financial companies that have a normalized LPTB greater than the threshold which we compare against the market-cap weighted portfolio.

this end, we create a simple trading strategy where we invest in the subset of the financial securities whose LPTB is greater than some threshold. In Figure 12f, we see this simple approach outperforms a market-cap weighted strategy; prior work (Monge et al., 2023) has found that growth stocks have been outperforming value stocks. Importantly, we note that the goal of this is to not introduce a new signal but to show how causal discovery can be a tool in searching for statistically significant relations which, if the assumptions are true for the data, can be interpreted as causal.

Evaluating Assumptions Similar to previous case studies, we see statistically significant lagged relationships and statistically significant nonstationarity, implying we would not be able to trust the results of an algorithm that cannot handle these types of properties. For our linearity test, we test the linearity of the normalized LPTB and returns conditional on all other parent nodes of returns and find the p-value be to 0.0 (i.e., we can reject the null hypothesis that the relationship is linear).

6 Conclusion and Future Work

In this paper, we developed a novel algorithm, Constraint-based Causal Discovery from Heterogeneous Time-Series (CD-NOTS), which is nonparametric (meaning it can capture both linear and non-linear causal relations), can handle nonstationarity, and is able to detect both lagged and contemporaneous causalities. We showed using simulated data that, for effective use of CD-NOTS, one should use ParCorr for low data regimes (less than 200 data points) and KCIT or RCoT for higher data regimes.

We applied CD-NOTS to Fama-French factors and Apple's returns, where we found that for small time ranges (approximately seven years), Apple's returns are stationary conditioning on the factors. This justifies the usage of Fama-French factors as a method for returns attribution. However, over longer time ranges (approximately two decades), we find Apple's returns are nonstationary even after accounting for the Fama-French factors, implying that there is a need to continually update the factor exposures of Apple through time.

Further, we found the causal relationship of economic factors (specifically unemployment, CPI, and PPI) tends to be similar across countries, especially for those geographically and economically similar. This finding implies that we should be able to create stronger models in economics by training a single model across different countries.

Finally, we apply CD-NOTS to company financials and stock returns and find a causal relationship between the Price-to-Book ratio and future returns of financial companies, justifying the common usage of Price-to-Book ratio for investing.

Besides our contribution in developing an algorithm that can handle nonstationarity, non-linear relationships, lagged and contemporaneous causalities, we showed the importance of these assumptions in finance. Specifically, many of the assumptions made in prior work (notably linearity, no lagged relationships, and stationarity) often do not hold in finance and thus should be used with caution. Hence, through weakening the assumptions required for our causal discovery algorithm to be used, we can trust more the causal relationships found by CD-NOTS and continue to explore the causal relationships in other financial datasets.

References

Baba, K., R. Shibata, and M. Sibuya (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics 46*(4), 657–664.

Bodenham, D. and N. Adams (2015, 06). A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing 26*.

Buckley, M. and G. Eagleson (1988). An approximation to the distribution of quadratic forms in normal random variables. *Australian Journal of Statistics 30A*(1), 150–159.

Buntine, W. (1991). Theory refinement on bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'91, San Francisco, CA, USA, pp. 52–60. Morgan Kaufmann Publishers Inc.

Bühlmann, P., M. Kalisch, and M. Maathuis (2009, 06). Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika 97*.

Chickering, D. M. (2003, mar). Optimal structure identification with greedy search. *3*(null), 507–554.

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

Entner, D. and P. Hoyer (2010, 09). On causal discovery from time series data using fci. *Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010*.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Gendron, G., M. Witbrock, and G. Dobbie (2023). A survey of methods, challenges and perspectives in causality.

Ghassami, A., N. Kiyavash, B. Huang, and K. Zhang (2018). Multi-domain causal structure learning in linear systems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Glymour, C., K. Zhang, and P. Spirtes (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics 10*.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Granger, C. W. J. (2001). *Investigating Causal Relations by Econometric Models and Cross-Spectral Methods*, pp. 31–47. USA: Harvard University Press.

Gretton, A., K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola (2008, September). A kernel statistical test of independence. In *Advances in neural information processing systems 20*, Red Hook, NY, USA, pp. 585–592. Max-Planck-Gesellschaft: Curran.

Hall, P. (1983). Chi Squared Approximations to the Distribution of a Sum of Independent Random Variables. *The Annals of Probability 11*(4), 1028 – 1036.

Huang, B., K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf (2020). Causal discovery from heterogeneous/nonstationary data with independent changes.

Huang, B., K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf (2017, November). Behind distribution shift: Mining driving forces of changes and causal arrows. In *IEEE 17th International Conference on Data Mining (ICDM)*, pp. 913–918.

Hyvärinen, A., K. Zhang, S. Shimizu, and P. O. Hoyer (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research 11*(56), 1709–1731.

Lam, W.-Y., B. Andrews, and J. Ramsey (2022, 01–05 Aug). Greedy relaxations of the sparsest permutation algorithm. In J. Cussens and K. Zhang (Eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, Volume 180 of *Proceedings of Machine Learning Research*, pp. 1052–1062. PMLR.

López de Prado, M. M. (2023). *Causal Factor Investing: Can Factor Investing Become Scientific?* Elements in Quantitative Finance. Cambridge University Press.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, San Francisco, CA, USA, pp. 403–410. Morgan Kaufmann Publishers Inc.

Monge, M., A. Lazcano, and J. L. Parada (2023). Growth vs value investing: Persistence and time trend before and after covid-19. *Research in International Business and Finance 65*, 101984.

Moraffah, R., P. Sheth, M. Karami, A. Bhattacharya, Q. Wang, A. Tahir, A. Raglin, and H. Liu (2021). Causal inference for time series analysis: Problems, methods and evaluation.

Nogueira, A. R., A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama (2022). Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery 12*(2), e1449.

Palus, M., V. Komarek, Z. Hrncír, and K. Sterbová (2001, 05). Synchronization as adjustment of information rates: Detection from bivariate time series. *Physical review. E, Statistical, nonlinear, and soft matter physics 63*, 046211.

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

Pearl, J. (2009a). *Causality* (2 ed.). Cambridge University Press.

Pearl, J. (2009b). *Causality: Models, Reasoning and Inference* (2nd ed.). USA: Cambridge University Press.

Pearl, J. (2014). The deductive approach to causal inference. *Journal of Causal Inference 2*(2), 115–129.

Peters, J., P. Bühlmann, and N. Meinshausen (2015). Causal inference using invariant prediction: identification and confidence intervals.

Peters, J., J. M. Mooij, D. Janzing, and B. Schölkopf (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research 15*(58), 2009–2053.

Ramsey, J. and B. Andrews (2017, 09). A comparison of public causal search packages on linear, gaussian data with no latent variables.

Ramsey, J., J. Zhang, and P. L. Spirtes (2012). Adjacency-faithfulness and conservative causal inference.

Rohekar, R. Y., S. Nisimov, Y. Gurwicz, and G. Novik (2021). Iterative causal discovery in the possible presence of latent confounders and selection bias. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*, Volume 34, pp. 2454–2465. Curran Associates, Inc.

Rubin, D. (1974, 10). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*.

Runge, J. (2017). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information.

Runge, J., P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances 5*(11), eaau4996.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 2*(6), 110–114.

Schreiber, T. (2000, July). Measuring information transfer. *Physical Review Letters 85*(2), 461–464.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461 – 464.

Schölkopf, B., F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio (2021). Towards causal representation learning.

Shimizu, S., P. O. Hoyer, A. Hyvärinen, and A. Kerminen (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research 7*(72), 2003–2030.

Shimizu, S., T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research 12*(33), 1225–1248.

Silander, T. and P. Myllymäki (2012, 06). A simple approach for finding the globally optimal bayesian network structure. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*.

Spirtes, P., C. Glymour, and R. Scheines (2001, 01). *Causation, Prediction, and Search*. The MIT Press.

Spirtes, P., C. Glymour, R. Scheines, S. A. Kauffman, V. Aimale, and F. C. Wimberly (2000). Constructing bayesian network models of gene expression networks from microarray data.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed (1990). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science 5*(4), 465 – 472.

Strobl, E. V., K. Zhang, and S. Visweswaran (2017). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery.

Tsamardinos, I., L. Brown, and C. Aliferis (2006, 10). The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning 65*, 31–78.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika 29*(3/4), 350–362.

Zhang, K., B. Huang, J. Zhang, C. Glymour, and B. Schölkopf (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 1347–1353.

Zhang, K. and A. Hyvärinen (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, Arlington, Virginia, USA, pp. 647–655. AUAI Press.

Zhang, K. and A. Hyvarinen (2012). On the identifiability of the post-nonlinear causal model.

Zhang, K. and A. Hyvärinen (2010, 12 Dec). Distinguishing causes from effects using nonlinear acyclic causal models. In I. Guyon, D. Janzing, and B. Schölkopf (Eds.), *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, Volume 6 of *Proceedings of Machine Learning Research*, Whistler, Canada, pp. 157–164. PMLR.

Zhang, K., J. Peters, D. Janzing, and B. Schoelkopf (2012). Kernel-based conditional independence test and application in causal discovery.

A Conditional Independence Tests Hyperparameters

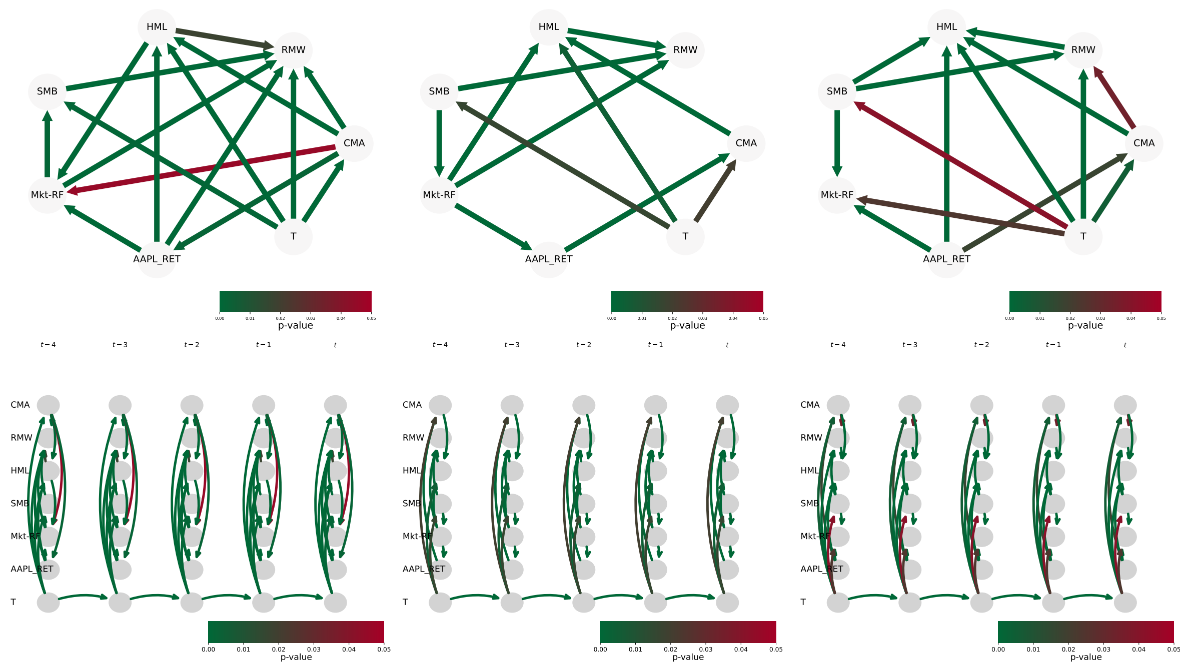
For all conditional independence tests, we use a threshold of 0.05 for the p-value.

KCIT For KCIT, we use the RBF kernel where the bandwidth is set to the median distance between the first five hundred data points.

RCoT For RCoT, similar to KCIT, we use the RBF kernel where the bandwidth is set to the median distance between the first five hundred data points. Further, we use a fixed number of Fourier features, specifically 25 for Z and 5 for $\ddot{X}$, X, and Y, similar to Strobl et al. (2017).

B Fama-French Factors Case Study with RCoT

We repeat the plots from our first case study (Section 5.1) using RCoT as opposed to KCIT. One empirical finding was that, due to the randomness of RCoT, many times there could be an inconsistency between the discovered graph and the CI tests, i.e., the graph shows conditional independences that are not found by the CI tests. We leave it to future work to find how to account for these forms on inconsistencies.



(a) Jan 2000 - Dec 2007 (b) Jan 2008 - Dec 2015 (c) Jan 2016 - Dec 2022

Figure 13: Results of CD-NOTS RCoT run on three non-overlapping time periods.

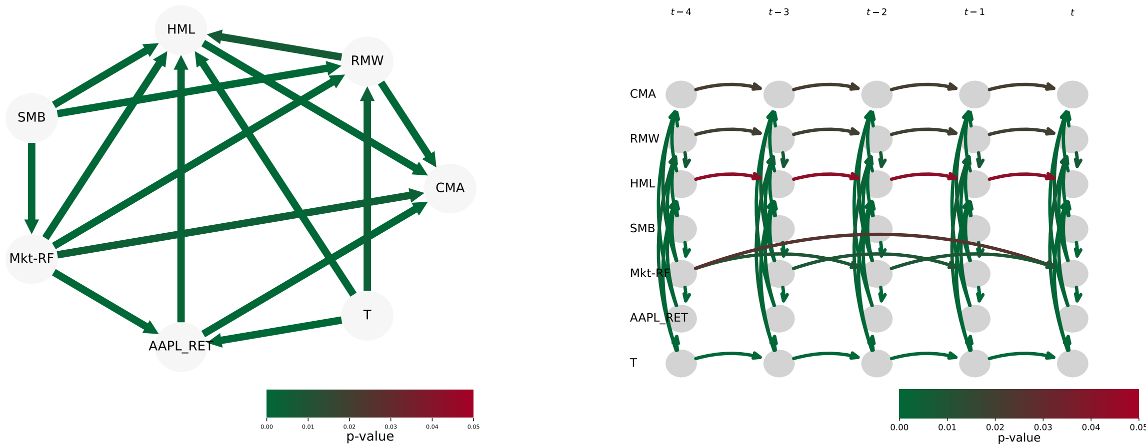
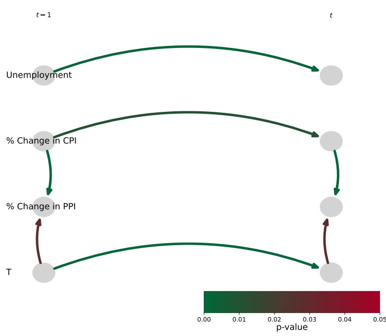
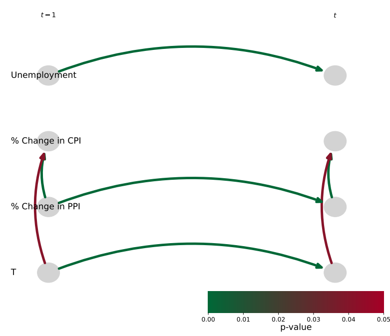


Figure 14: Results of CD-NOTS RCoT run from beginning of 2000 to end of 2022.

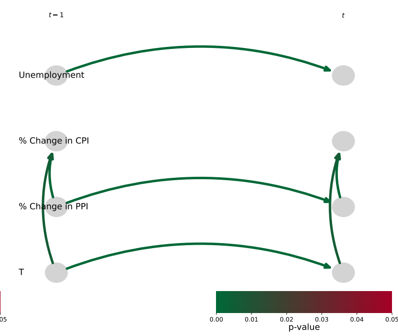
C Economic Data Case Study with KCIT and RCoT



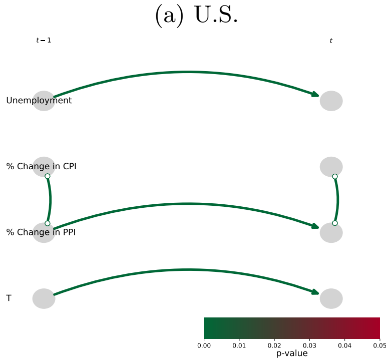
(a) U.S.



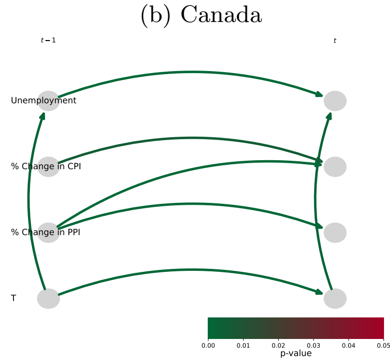
(b) Canada



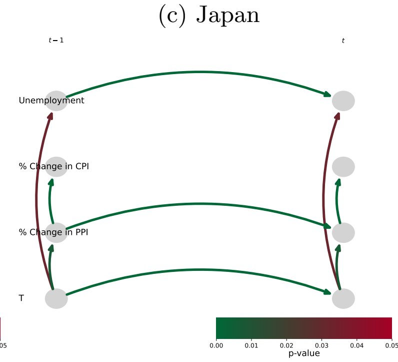
(c) Japan



(d) France

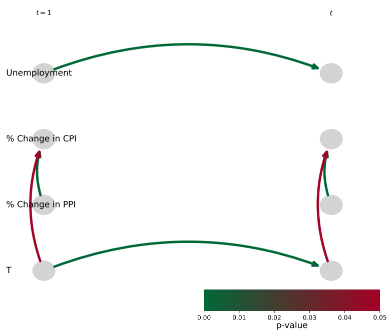


(e) United Kingdom

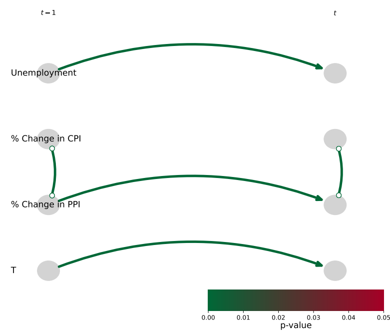


(f) Italy

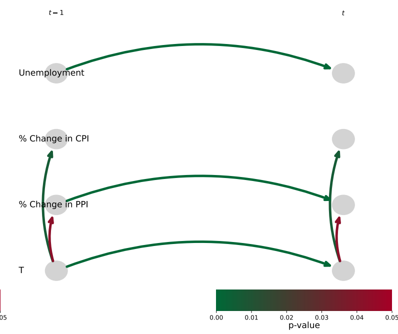
Figure 15: Results of CD-NOTS KCIT run on six countries' economic variables.



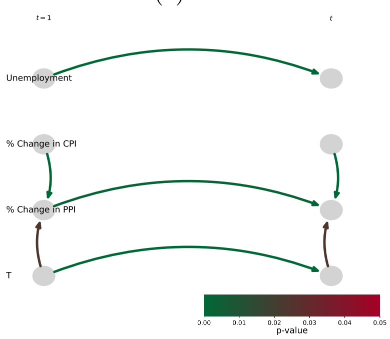
(a) U.S.



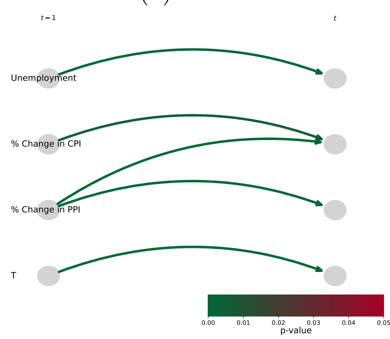
(b) Canada



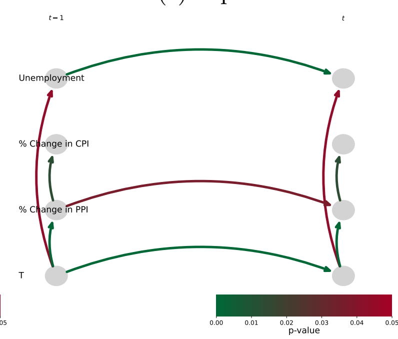
(c) Japan



(d) France



(e) United Kingdom



(f) Italy

Figure 16: Results of CD-NOTS RCoT run on six countries' economic variables.

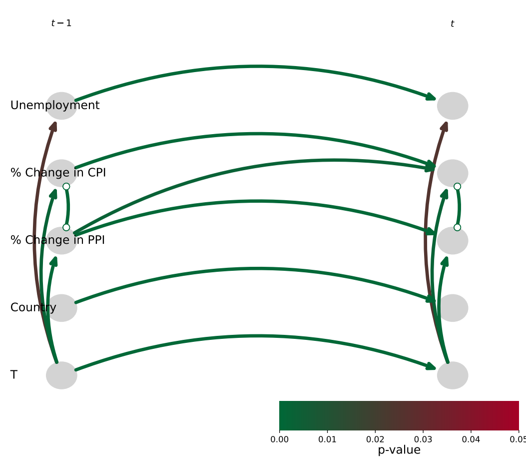


Figure 17: Results of CD-NOTS RCoT run on the economic variables of multiple countries.

D Detailed Simulation Results

D.1 Different CI Tests

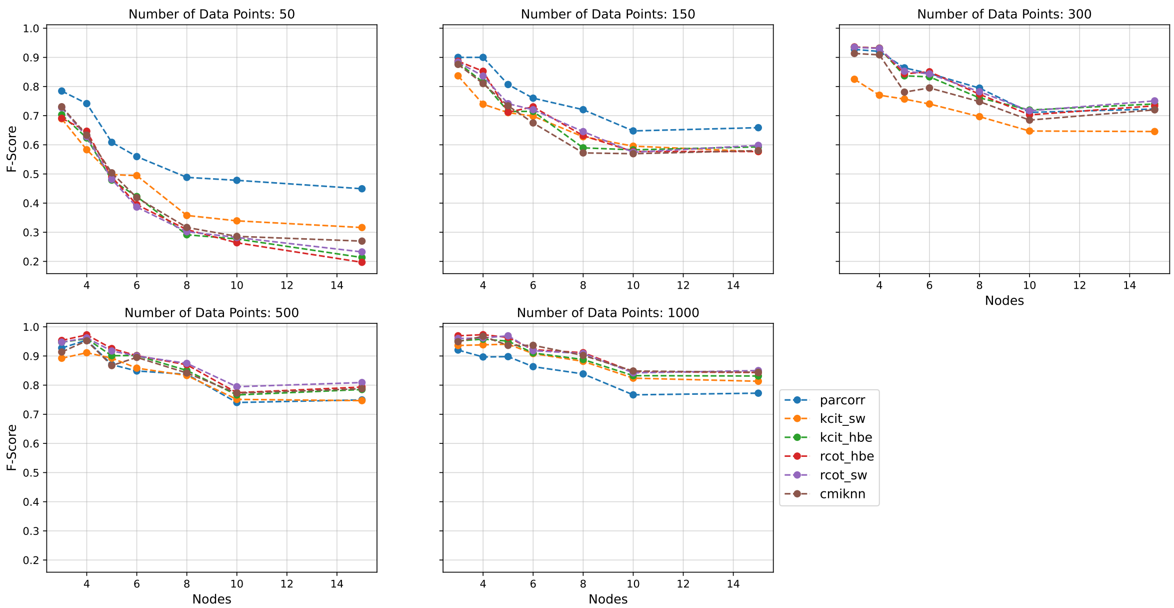


Figure 18: F-score evaluation for CD-NOTS with different CI tests, tested out on many different simulated datasets.

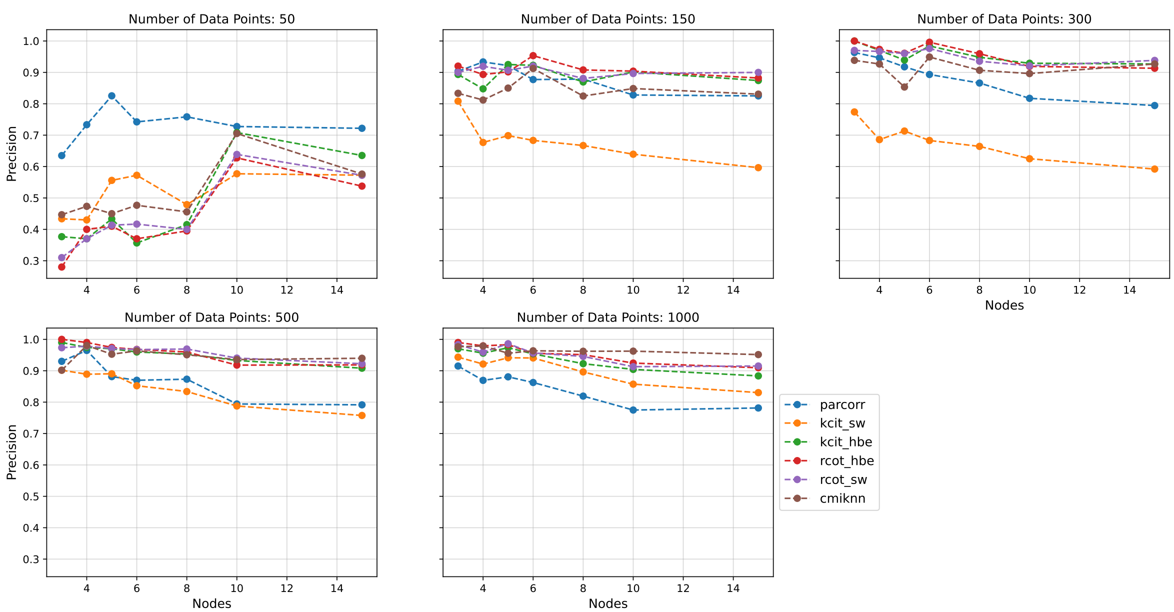


Figure 19: Precision evaluation for CD-NOTS with different CI tests, tested out on many different simulated datasets.

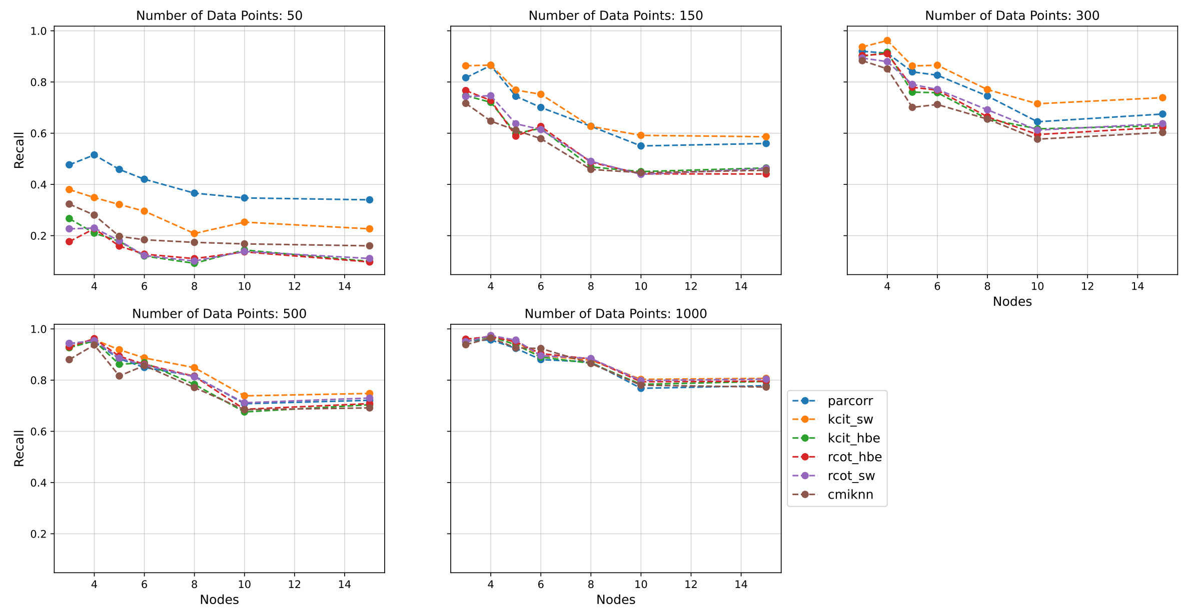


Figure 20: Recall evaluation for CD-NOTS with different CI tests, tested out on many different simulated datasets.

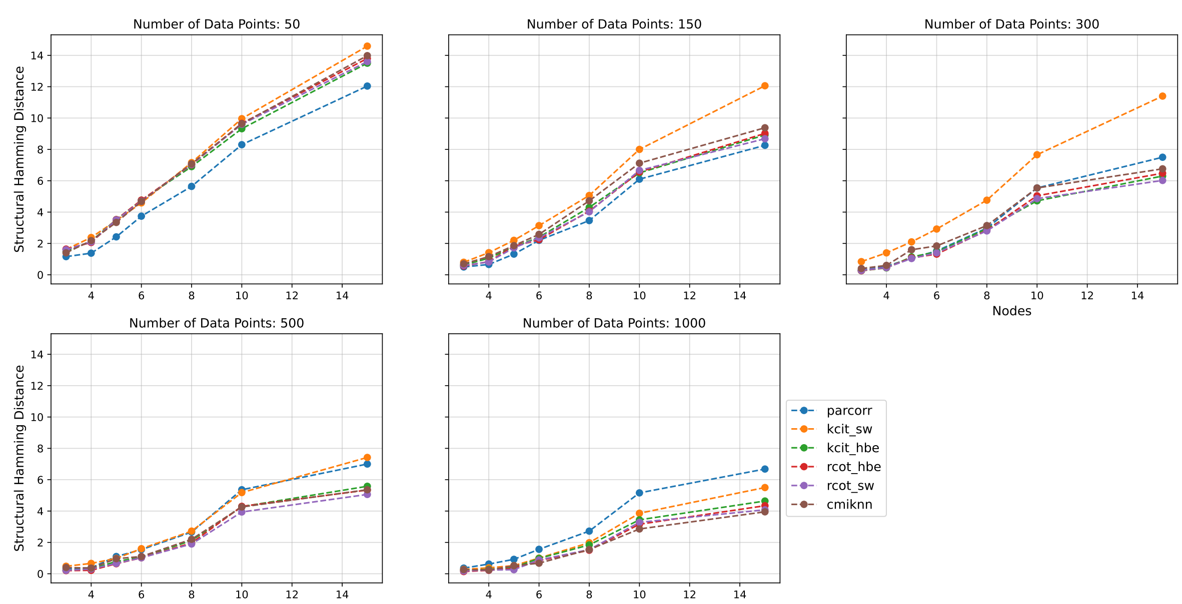


Figure 21: Structural Hamming Distance evaluation for CD-NOTS with different CI tests, tested out on many different simulated datasets.

D.2 Comparison against PCMCI

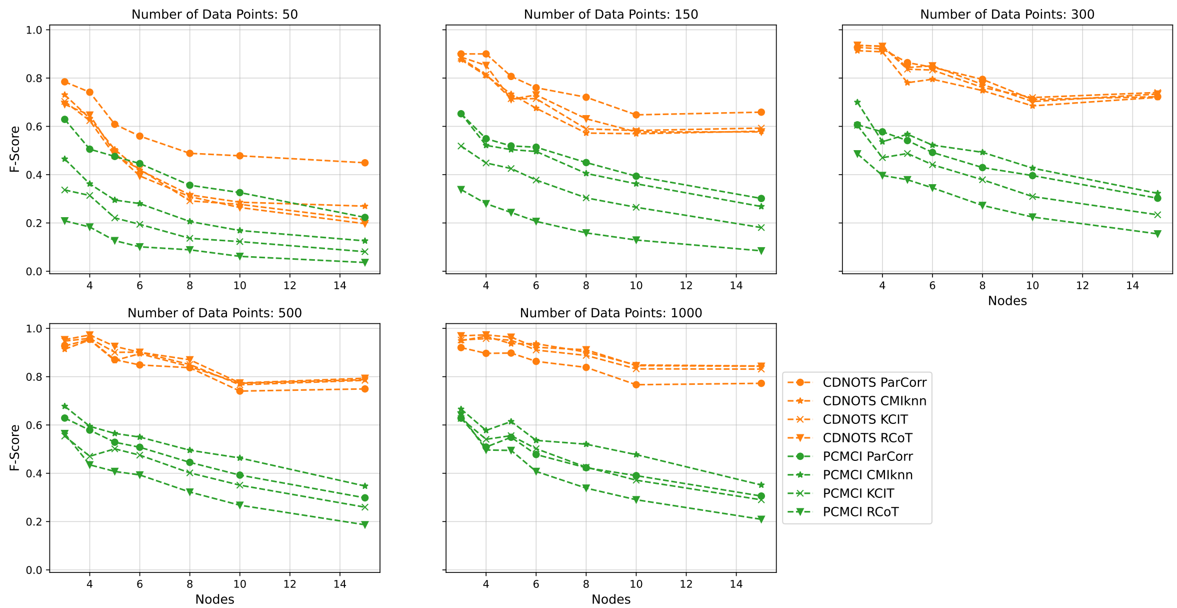


Figure 22: F-score evaluation, comparing CD-NOTS and PCMCI, tested out on many different simulated datasets.

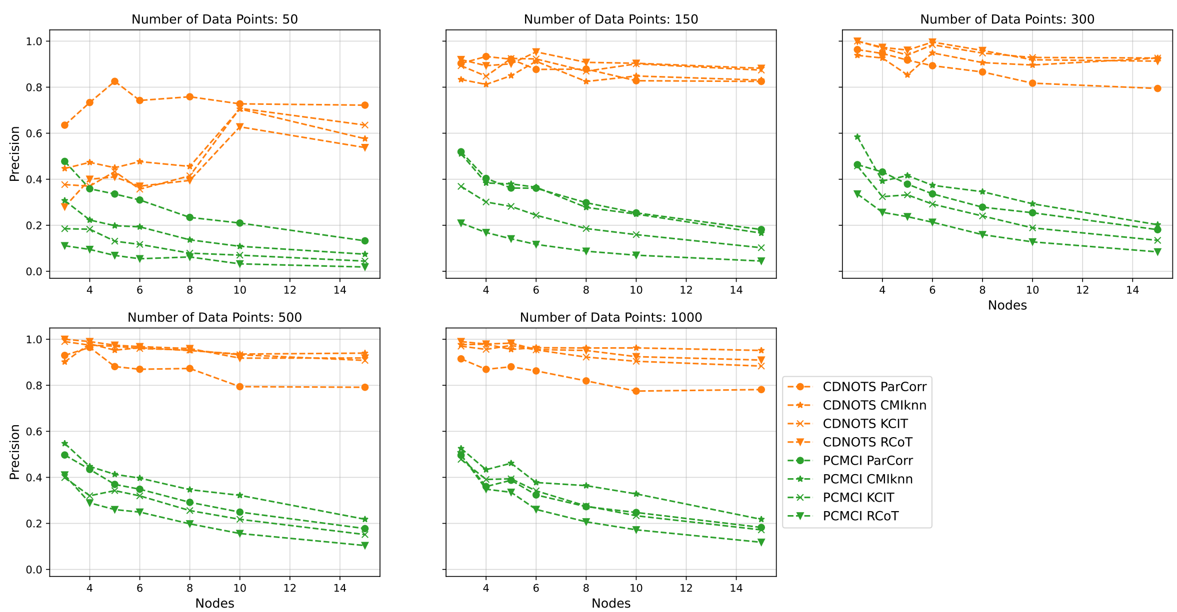


Figure 23: Precision evaluation, comparing CD-NOTS and PCMCI, tested out on many different simulated datasets.

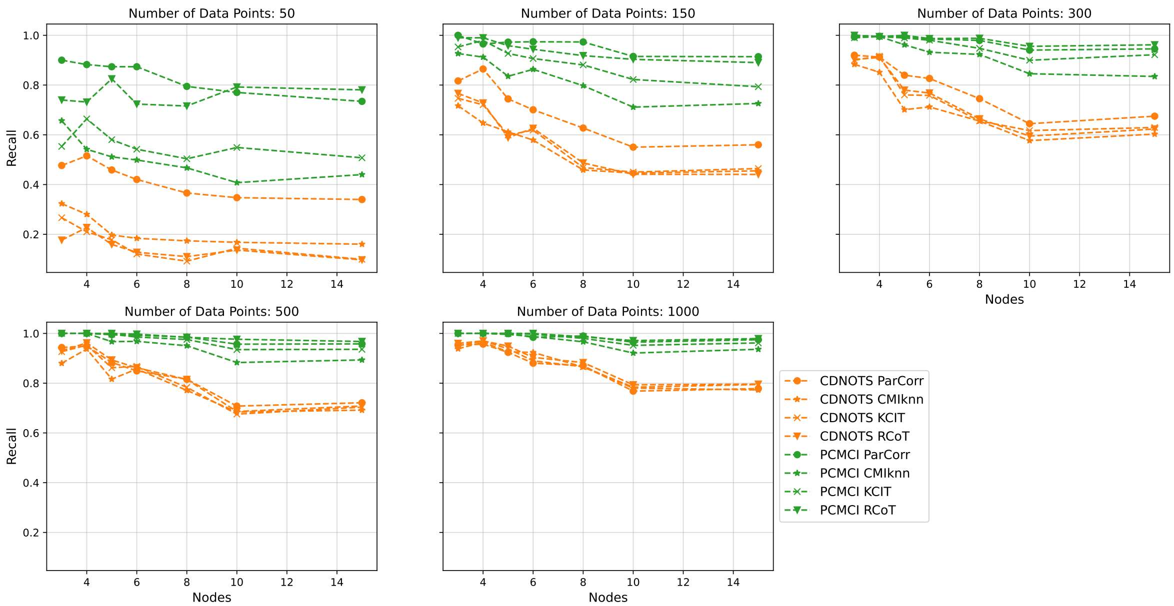


Figure 24: Recall evaluation, comparing CD-NOTS and PCMCI, tested out on many different simulated datasets.

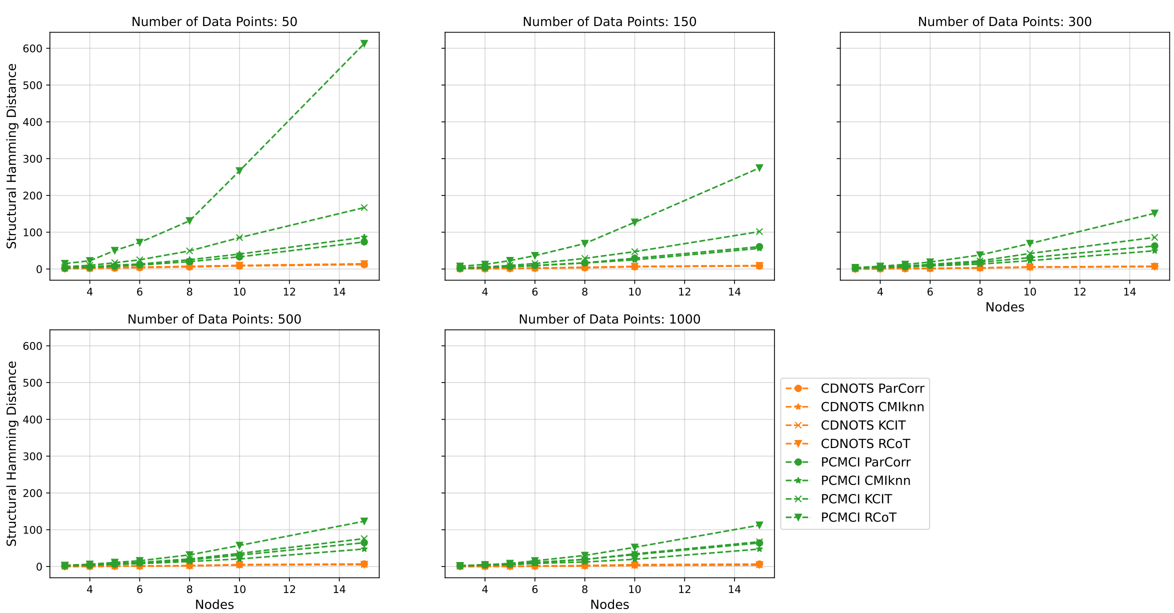


Figure 25: Structural Hamming Distance evaluation, comparing CD-NOTS and PCMCI, tested out on many different simulated datasets.