Discovering Time-invariant Causal Structure from Temporal Data

Saima Absar
University of Arkansas
sa059@uark.edu

Lu Zhang
University of Arkansas
lz006@uark.edu

Abstract

Discovering causal structure from temporal data is an important problem in many fields in science. Existing methods usually suffer from several limitations such as assuming linear dependencies among features, limiting to discrete time series, and/or assuming stationarity, i.e., causal dependencies are repeated with the same time lag and strength at all time points. In this paper, we propose an algorithm called the μ-PC that addresses these limitations. It is based on the theory of μ-separation and extends the well-known PC algorithm to the time domain. To be applicable to both discrete and continuous time series, we develop a conditional independence testing technique for time series by leveraging the Recurrent Marked Temporal Point Process (RMTPP) model. Experiments using both synthetic and real-world datasets demonstrate the effectiveness of the proposed algorithm.

CCS Concepts

• **Mathematics of computing → Causal networks**.

Keywords

Causal discovery; temporal data; μ-separation; directed graph; recurrent neural network

ACM Reference Format:

Saima Absar and Lu Zhang. 2021. Discovering Time-invariant Causal Structure from Temporal Data. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482086

1 Introduction

Discovering causal relationships among a set of variables is a fundamental problem in many fields in science. The gold standard of causal discovery is randomized controlled trials, which is usually not feasible due to ethical issues or unacceptably high costs. In the past decades, many research works have been conducted on discovering causal relationships from observational data. However, the majority of these works are based on static settings [11]. In many applications, the data is temporal in nature. Incorporating temporal information is an important extension to the causal discovery field.

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482086

The classic definition of temporal causality is the Granger causality [12]. It states that one time series is the cause of the other time series if the former's past values can predict the latter's future values. Modern theories of causal discovery and inference often rely on a more robust mathematical framework called structural causal models [21] and are conceptually visualized by a graphical model. Under the causal Markov and faithfulness assumptions, a causal graph encodes all conditional relations contained in the data, by using graphical criteria such as d-separation. Since in practice, only statistical properties of observational data can be analyzed, constraint-based search algorithms (e.g., the PC algorithm) try to search for causal graphs that most closely entail conditional independence relations held in the data [26]. In [23], Runge proposes the PCMCI algorithm that is extended from the classic PC algorithm. The PCMCI builds time-lagged causal graphs from time series datasets, where each variable in a time series is represented as one node in the graph. However, it applies to discrete time series only and it assumes stationarity, i.e., causal dependencies are repeated with the same time lag at all time points, which may not hold in many applications.

In this paper, we propose a method to build a time-invariant causal graph where each node in the graph represents a whole time series. Different from the static setting where a causal graph is a directed acyclic graph, a time-invariant causal graph for time series is a directed graph where circles and self-loops may exist. Previous works on building directed graphs for time series are mainly extended from the Granger causality to multivariate processes using linear or nonlinear autoregression models [10, 22]. In our work, we follow the logic of constraint-based search algorithms to find directed graphs that agree most closely with conditional independence relations held in time series data. We leverage the μ-separation which is an extension of d-separation to the time-dependent domain [19]. The theory of μ-separation provides a formal graphical representation of conditional independence relations in time series with rigorous analysis of the equivalence class of graphs. It does not assume stationarity in causal dependencies, i.e., the time lags and strength of causal relationships may vary with time. In addition, different from the Granger causality that always assumes there is a dependence of each process on its own past, the μ-separation can also detect if there exists confounding for a time series with itself.

Based on the μ-separation, we develop the μ-PC algorithm, a constraint-based algorithm for building directed graphs from time series data following similar logic to the PC algorithm. Inspired by [3], we modify the classic PC algorithm as two phases, where Phase-1 is a pre-processing stage that selects the candidate parental set for each node, and Phase-2 continues to prune the graph based on the μ-separation. To handle both discrete and continuous time series, we leverage the Recurrent Marked Temporal Point Process (RMTPP) model proposed in [9] for Conditional Independence (CI) testing since the marked temporal point process is a commonly used mathematical framework for modeling dynamic events with both event timings and markers. We apply the likelihood ratio test to examine if the latter is significantly better than the former. We conduct experiments using both synthetic and real-world climate datasets. The results show that the proposed algorithm is capable of recovering directed graphs with circles and self-loops, and outperform the PCMCI when the stationary assumption is not satisfied.

Related Work. The classic causal notion for time series is the Granger causality, based on which various methods were proposed that differ in how to measure the predictability (e.g., [2, 16, 17, 27]). Modern causal inference methods extend the Granger causality based on structural causal models and associated graphical representations [3, 4, 8, 18]. A summary of relevant background and a comprehensive review of different approaches to causal discovery is provided in [20, 29]. Machine learning and deep learning based methods have been proposed in recent years to handle complex data. [6] proposes to replace regression models with classifiers by formulating a feature representation that could capture the causal relationship. Tank et al. [28] claims to extract the nonlinear Granger causality in time series using regularized neural network models like Multi-Layer Perceptron (MLP) and Recurrent Neural Network (RNN). Recently in [30], the authors propose an encoder-decoder architecture for Neural Point Process (NPP) that can learn event inter-dependencies, followed by an attribution method to learn the Granger causality among multi-type event sequences. [1] utilizes generalized linear models for fitting the data and uses the minimum description length principle to determine causal directions. A similar approach based on mutual information has been proposed by Jangyodsuk et al. [13] where the authors claim that causality in time series can be determined using mutual information between an effect and a cause. Another line of work is the PCMCI algorithm proposed by Runge [23–25] which extends the classic PC algorithm to time series. However, it assumes stationarity which may not hold in practice. Moreover, most of the above works only apply to discrete time series. In this paper, we propose a theoretically sound constraint-based algorithm based on the μ-separation and RMTPP model so that we do not assume stationarity and the algorithm works for both discrete and continuous time series.

2 Preliminaries

Consider a set of time series $\mathcal{X} = (\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^n)$ where $\mathbf{X}^i = \{X_t^i : t \in [0, T]\}$. For continuous time series, t is a real value variable in a compact time interval; and for discrete time series, t is a discrete time index. We also use a set of indices $\mathbf{V} = \{1, 2, \ldots, n\}$ to refer to time series in $\mathcal{X}$ (as well as the nodes in a graph as shown later). Conditional (local) independence between two time series α, β given a subset of time series $\mathbf{C}$ is defined as follows [19].

Definition 1 (Conditional independence). *Let $\alpha, \beta \in \mathbf{V}$, $\mathbf{C} \subseteq \mathbf{V} \backslash \{\alpha\}$. We say that β is conditionally independent of α given $\mathbf{C}$ if for any time point t, the past of $\mathbf{X}^{\mathbf{C}}$ until time t gives us the same predictable information about X_t^{β} as the past of both $\mathbf{X}^{\alpha}$ and $\mathbf{X}^{\mathbf{C}}$ until time t, denoted by $CI(\alpha, \beta | \mathbf{C})$.*

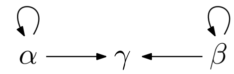

Figure 1: A directed graph.

Note that, the conditional independence relation for time series is asymmetric, i.e., $CI(\alpha, \beta|\mathbf{C})$ does not necessarily imply $CI(\beta, \alpha|\mathbf{C})$.

A causal graph for time series is a directed graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is a set of nodes and $\mathbf{E}$ is a set of directed edges. For each pair of connected nodes $\alpha \longrightarrow \beta$, it is called a loop if $\alpha = \beta$. For each node α, we denote its parents by $Pa(\alpha)$. A walk from α to β is a sequence of connected nodes including α and β as two end points, together with the edges connecting them. A route from α to β is a walk from α to β such that no node other than β occurs more than once, and β occurs at most twice. For example, in the directed graph shown in Fig. 1, $\alpha \longrightarrow \alpha \longrightarrow \gamma \longleftarrow \beta$ is a walk from α to β, and $\alpha \longrightarrow \gamma \longleftarrow \beta \longrightarrow \beta$ is a route from α to β. For any walk or route ω from α to β, if a node γ other than the end points is on ω and both edges have heads at it, i.e., $\longrightarrow \gamma \longleftarrow$, then γ is called a collider on ω; otherwise, it is a noncollider. Then, the μ-connecting walk/route and μ-separation are defined as follows [19].

Definition 2 (μ-connecting walk/route). *A walk/route ω is said to be μ-connecting from α to β given $\mathbf{C}$ if $\alpha \notin \mathbf{C}$, every collider on ω is in $\mathbf{C} \cup An(\mathbf{C})$, no noncollider is in $\mathbf{C}$, and the last edge has a head at β. When a walk/route is not μ-connecting given $\mathbf{C}$, it is said to be blocked by $\mathbf{C}$*

Definition 3 (μ-separation). *Let $\alpha, \beta \in \mathbf{V}$, $\mathbf{C} \subseteq \mathbf{V}$. We say that β is μ-separated from α given $\mathbf{C}$ if there is no μ-connecting route from α to β given $\mathbf{C}$.*

3 μ-PC Algorithm

We propose a constraint-based search algorithm called the μ-PC algorithm for time series. The target is to build a directed graph such that all μ-separations holding in the graph follow corresponding conditional independence relations in the data. We make the Causal Markov and Faithfulness assumptions which are routinely employed in independence-based causal discovery. Inspired by [3], our algorithm consists of two phases for improving efficiency and strengthening reliability. The pseudocode of the μ-PC Algorithm is shown in Algorithm 1.

Phase-1 is to pre-process the data to select the candidate parental set for each node. The algorithm starts with an empty parent set $Pa(\beta)$ for each node β. It performs conditional independence testing between the current node β and every node α in the graph, including itself (to check for self-loops). Line 8 calls a subroutine $score(\cdot)$ to conduct this conditional independence testing and returns the likelihood ratio test (LRT) score which will be discussed in the next section. For each node in the graph that has not been already added to $Pa(\beta)$, the LRT score is computed and stored in a set $\mathbf{U}$ along with its corresponding node name if the score is greater than zero. Afterward, the node with the highest score in $\mathbf{U}$ is added to $Pa(\beta)$. Given the most recently updated parental set, conditional independence testing is conducted for each remaining node again. The entire process is repeated until there is no node left to be added to $Pa(\beta)$ that has a score greater than zero.

In Phase-2, the algorithm proceeds with the graph formed with the parental sets in Phase-1. It then prunes the graph according to the μ-separation criterion: for $\alpha, \beta \in \mathbf{V}$, if there exists $\mathbf{C} \subseteq \mathbf{V} \backslash \{\alpha\}$ such that $CI(\alpha, \beta|\mathbf{C})$ holds, then there is no edge pointing from α to β in the graph. We can further restrict $\mathbf{C}$ to the sets such that there exists a μ-connecting route from α to β that is blocked by $\mathbf{C}$. The algorithm starts from all pairs of time series α, β (α can be same as β) with an empty condition set, i.e., $|\mathbf{C}| = 0$. If β is independent of α, it removes the edge pointing from α to β. Then, it proceeds to condition sets with one element except α, i.e., $|\mathbf{C}| = 1$, and continues to do so until $|\mathbf{C}| = n$. The algorithm conducts the test only when currently there exists a μ-connecting route from α to β blocked by $\mathbf{C}$. Different from the Granger causality that always assumes self-dependence for each time series, for each time series, the μ-PC Algorithm always examines whether it is self-independent, and if not, whether such dependence could be μ-separated by other time series. Finally, the theory of μ-separation ensures that every Markov equivalent class is a singleton that can be uniquely identified from the data [19].

Algorithm 1: μ-PC Algorithm

```
Phase-1
foreach β ∈ V do
    Pa(β) = ∅ ;
    repeat
        U = ∅ ;
        foreach α ∈ V such that α ∉ U do
            if score(α, β, U) > 0 then
                U ∪ score(α, β, U);
        γ = arg max(U);
        Pa(β) = Pa(β) ∪ {γ} ;
    until |U| > 0;
Phase-2
Start from the graph of Phase-1;
for i = 0 to n − 1 do
    foreach C ⊆ V, |C| = i do
        foreach α ∈ V such that α ∉ C do
            foreach β ∈ V such that α → β do
                if there exists a μ-connecting route from α to β that
                  is blocked by C then
                    if score(α, β, C) = 0 then
                        Remove edge from α to β ;
```

For the time complexity, the number of conditional independence tests required by the algorithm is $O(|\mathbf{V}|)^3$ in Phase-1 and exponential to the number of nodes in the worst case in Phase-2 similarly to the PC algorithm. For the latter, many techniques have been proposed for improving efficiency (e.g., [7, 14]). In addition, in practice, conditional independence testing is conducted only when there exists a corresponding μ-connecting route. A subroutine for finding all μ-connecting routes based on the depth first search has a $O(|\mathbf{V}| + |\mathbf{E}|)$ time complexity.

4 RMTPP-based CI Testing

The Conditional Independence (CI) testing for time series is the key component of the μ-PC algorithm. In order to handle both continuous and discrete time series, we leverage the Recurrent Marked Temporal Point Process (RMTPP) [9], which is a unified model capable of modeling general nonlinear dependency among time series for both variables values and timings. The RMTPP model is an RNN that takes the history of the time series as inputs to predict both the value and timing of the next observation. The hidden state $\mathbf{h}$ learns a representation of the influence from the history of the time series and is updated each time after reading in the next observation. Given the learned representation $\mathbf{h}$, the likelihood that the next observation will occur at time t^* relative to the timing of the latest observation t', denoted by $f(t^*|\mathbf{h})$, as well as the likelihood that the marker value of the next observation is x^*, denoted by $P(x^*|\mathbf{h})$, are computed. Consequently, the RMTPP model is trained by maximizing the log-likelihood over the entire time series dataset.

For CI testing, we make use of the RMTPP model to capture the predictable information contained in the time series. According to the definition of conditional independence, given any triple $\alpha, \beta, \mathbf{C}$, we fit two RMTPP models, referred to as m_0 and m_1. Conceptually, in m_0 we feed time series in $\mathbf{X}^{\mathbf{C}}$ into the model as inputs and predict time series $\mathbf{X}^{\beta}$; while in m_1, we use both $\mathbf{X}^{\alpha}$ and $\mathbf{X}^{\mathbf{C}}$ to predict $\mathbf{X}^{\beta}$. If the performance of the two models is the same, it means that $\mathbf{X}^{\mathbf{C}}$ contains the same predictable information about $\mathbf{X}^{\beta}$ as $\mathbf{X}^{\alpha}$ and $\mathbf{X}^{\mathbf{C}}$. In the implementation, we actually use the same inputs $\mathbf{X}^{\alpha}$ and $\mathbf{X}^{\mathbf{C}}$ for both models but restrict the parameters associated with $\mathbf{X}^{\alpha}$ in m_0 to be zeros. This allows us to employ the likelihood ratio test for measuring predictability, which is a statistical test of the goodness-of-fit between two models, a more complex one with more parameters and a simpler one with fewer parameters.

Different from the original RMTPP model where the inputs and outputs are the same time series, in our context the inputs and outputs can be different time series. To deal with this change, we align all the time series on the same temporal dimension and feed the observations into the model according to the temporal order. If the current observation is for input, i.e., X_t^{α} or $X_t^{\mathbf{C}}$, we update the hidden state $\mathbf{h}$; otherwise, we compute the likelihood of X_t^{β} for the cost function. Symbolically, for the triple $\alpha, \beta, \mathbf{C}$, the unified formula for updating $\mathbf{h}$ at time t is given by

$$\mathbf{h}_{new} = \max\left\{ \sum_{i \in C \cup \{\alpha\}} I^i \left(\mathbf{W}^i \mathbf{X}_t^i + \mathbf{V}^i t \right) + \mathbf{W}^h \mathbf{h} + \mathbf{b}^h, 0 \right\}, \quad (1)$$

where $\mathbf{X}_t^i$ is the embedding of X_t^i, and $\mathbf{W}^i, \mathbf{V}^i$ are parameters associated with the input marker and timing respectively. I^i is an indicator parameter that is restricted to zero if $i = \alpha$ in m_0. Thus, we consider there is one parameter difference between m_0 and m_1.

Denoting by $t = \{t_1, \ldots, t_j, \ldots, t_m\}$ the timings of $\mathbf{X}^{\beta}$, the cost function is given by

$$l = \sum_{j=1}^{m} \left(\log P(X_{t_j}^{\beta} | \mathbf{h}_{j^-}) + \log f(t_j | \mathbf{h}_{j^-}) \right), \quad (2)$$

where $\mathbf{h}_{j^-}$ denotes the latest updated hidden state prior to t_j. Finally, the network architecture of our CI testing model is shown in Fig. 2.

Denoted by l_0 and l_1 are the maximized log-likelihood obtained by m_0 and m_1 respectively. The likelihood ratio test is conducted upon the ratio $\lambda = l_0/l_1$. The null hypothesis is that the more complex model m_1 is not significantly better than the simpler model m_0. Thus, the conditional independence relation holds if the null

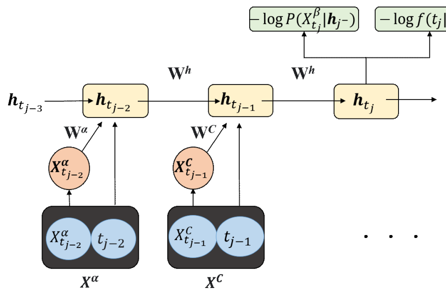

Figure 2: Network architecture of RMTPP-based CI testing. For each triple $\alpha, \beta, \mathrm{C}$, we input the timings and values of X^{α} and X^{β} into the RMTPP for updating hidden state h, while treating X^{β} as the output for computing the cost.

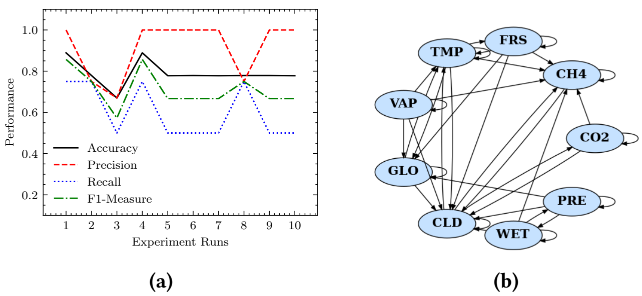

Figure 3: (a) Results of μ-PC on continuous time series. (b) Causal graph of climate data in Montana.

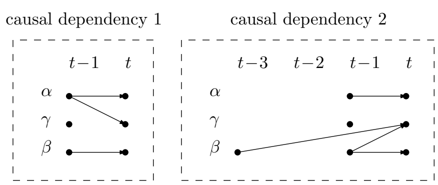

Figure 4: Time-lagged causal graphs for data generation. The value of X^{γ} at each time is generated by randomly selecting either one of the causal dependencies.

hypothesis is not rejected. The subroutine $score(\alpha, \beta, \mathrm{C})$ is hence computed as $score = \max((\chi^2 - \tau), 0)$ where $\chi^2 = -2 \log \lambda$ and τ is a threshold obtained from the 1% percentile point of Chi-distribution with 1 degree of freedom.

Remark. For discrete time series, there are no timings and the values are simply indexed in time order. In this case, we can use a simplified version of the model by removing parameter matrices and log-likelihood terms associated with the time from Eq. (1) and Eq. (2) respectively. The CI testing procedure remains the same.

5 Experiments

We conduct experiments on synthetic and real-world datasets to evaluate the performance of the μ-PC algorithm. We compare our method with PCMCI [23], ITGH [1], and CUTE [5]. For PCMCI, we convert its time-lagged causal graph to a directed graph such that if there is an edge from one variable to another with any lag in the time-lagged causal graph, then there is an edge between corresponding nodes in the directed graph.

Synthetic Data. We utilize the directed graph with three nodes α, β, γ as shown in Fig. 1. Then, we define two mechanisms for

	Accuracy	Precision	Recall	F-Measure
μ-PC	0.900	0.895	0.900	0.888
ITGH	0.667	0.667	0.500	0.572
PCMCI	0.867	0.897	0.850	0.862
CUTE	0.355	0.350	0.525	0.420

Table 1: Results on discrete time series w/o stationarity.

generating both timings and marker values of each time series based on the graph. In Setting 1, to evaluate the μ-PC algorithm for continuous time series, we generate the timings as Hawkes processes and marker values with multinomial distribution. The results of the 10 runs are shown in Fig. 3(a). As we can see, in general, the μ-PC algorithm could correctly identify the causal dependencies as well as directions. However, the performance is not quite stable. This may be due to the randomness in the training process and the imperfect fitting to the observational distribution. Particularly, our current algorithm is based on the traditional RNN which is not able to keep track of long-term dependencies. We will study if adopting modern versions of RNN such as LSTM or GRU could improve the performance of our algorithm. In Setting 2, we generate discrete time series by generating marker values for each time index. To investigate the effect of non-stationarity causal dependencies, we consider two types of causal dependencies with different time-lags and strengths that are used randomly in the generation as shown in Fig. 4. The performance of our method and baselines is shown in Table 1. We see that our method significantly outperforms ITGH and CUTE, and slightly outperforms PCMCI, showing the improvement of our method relative to PCMCI in nonstationary settings.

Climate Data. Finally, we conduct experiments on the climate data used in [1, 15]. The dataset contains monthly measures of 9 features over 13 years taken in Montana including temperature (TMP), frost days (FRS), greenhouse gases including Methane (CH4), carbon dioxide (CO2), solar radiation including global extraterrestrial (GLO), precipitation (PRE), vapor (VAP), cloud cover (CLD), and wet days (WET). The causal graph built by our method is shown in Fig. 3(b). We see that our method can discover edges that have reasonable explanations in climate science. For example, the causal graph shows that the vapor does not directly influence the number of frost days but the effect is indirect through cloud cover and temperature. However, there are some edges that do not have clear explanations. Possible reasons for these edges include insufficient data and hidden variables.

6 Conclusions

In this paper, we proposed an algorithm called the μ-PC for building causal graphs for time series. The algorithm is based on the theory of μ-separation so it does not assume stationarity. We proposed an RMTPP-based CI testing technique so that the algorithm works for both discrete and continuous time series. Experiments using both synthetic and real-world datasets showed that the μ-PC algorithm could recover causal relationships from time series more accurately than existing methods. Our work established a general framework for designing constraint-based search algorithms for time series.

Acknowledgments

This work was supported in part by NSF 1920920 and 1946391.

References

[1] Sahar Behzadi, Benjamin Schelling, and Claudia Plant. 2020. ITGH: Information-Theoretic Granger Causal Inference on Heterogeneous Data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 742–755.

[2] David Bell, Jim Kay, and Jim Malley. 1996. A non-parametric approach to non-linear causality testing. *Economics Letters* 51, 1 (1996), 7–18.

[3] Debarun Bhattacharjya, Karthikeyan Shanmugam, Tian Gao, and Dharmashankar Subramanian. 2020. Structure Discovery in (Causal) Proximal Graphical Event Models. In *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*.

[4] Debarun Bhattacharjya, Dharmashankar Subramanian, and Tian Gao. 2018. Proximal graphical event models. *Advances in Neural Information Processing Systems* 31 (2018), 8136–8145.

[5] Kailash Budhathoki and Jilles Vreeken. 2018. Causal inference on event sequences. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 55–63.

[6] Yoichi Chikahara and Akinori Fujino. 2018. Causal Inference in Time Series via Supervised Learning.. In *IJCAI*. 2042–2048.

[7] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics* (2012), 294–321.

[8] Vanessa Didelez. 2008. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 1 (2008), 245–264.

[9] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1555–1564.

[10] Michael Eichler. 2012. Graphical modelling of multivariate time series. *Probability Theory and Related Fields* 153, 1-2 (2012), 233–268.

[11] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

[12] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[13] Piraporn Jangyodsuk, Dong-Jun Seo, and Jean Gao. 2014. Causal Graph Discovery For Hydrological Time Series Knowledge Discovery. (2014).

[14] Thuc Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. 2016. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM transactions on computational biology and bioinformatics* (2016).

[15] Yan Liu, Alexandru Niculescu-Mizil, Aurelie C Lozano, and Yong Lu. 2010. Learning temporal causal graphs for relational time-series analysis. In *ICML*.

[16] Helmut Lütkepohl. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

[17] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. 2008. Kernel-Granger causality and the analysis of dynamical networks. *Physical review E* 77, 5 (2008), 056215.

[18] Christopher Meek. 2014. Toward Learning Graphical and Causal Process Models.. In *CI@ UAI*. 43–48.

[19] Søren Wengel Mogensen, Niels Richard Hansen, et al. 2020. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics* 48, 1 (2020), 539–559.

[20] Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. 2021. Causal Inference for Time series Analysis: Problems, Methods and Evaluation. *arXiv preprint arXiv:2102.05829* (2021).

[21] Judea Pearl. 2009. *Causality*. Cambridge university press.

[22] Christopher J Quinn, Negar Kiyavash, and Todd P Coleman. 2015. Directed information graphs. *IEEE Transactions on information theory* 61, 12 (2015), 6887–6909.

[23] Jakob Runge. 2018. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28, 7 (2018), 075310.

[24] Jakob Runge. 2018. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 938–947.

[25] Jakob Runge. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. *arXiv preprint arXiv:2003.03685* (2020).

[26] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[27] Xiaohai Sun. 2008. Assessing nonlinear Granger causality from multivariate time series. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 440–455.

[28] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. 2018. Neural granger causality for nonlinear time series. *stat* 1050 (2018), 16.

[29] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. 2021. D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. *arXiv preprint arXiv:2103.02582* (2021).

[30] Wei Zhang, Thomas Panum, Somesh Jha, Prasad Chalasani, and David Page. 2020. Cause: Learning granger causality from event sequences using attribution methods. In *International Conference on Machine Learning*. PMLR, 11235–11245.