Causal Discovery from Temporal Data: An Overview and New Perspectives

CHANG GONG, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China and University of the Chinese Academy of Sciences, Beijing, China
CHUZHE ZHANG, School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore
DI YAO, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
JINGPING BI, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China
WENBIN LI, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China and University of the Chinese Academy of Sciences, Beijing, China
YONGJUN XU, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China

Temporal data, representing chronological observations of complex systems, has always been a typical data structure that can be widely generated by many domains, such as industry, finance, healthcare, and climatology. Analyzing the underlying structures, i.e., the causal relations, could be extremely valuable for various applications. Recently, causal discovery from temporal data has been considered as an interesting yet critical task and attracted much research attention. According to the nature and structure of temporal data, existing causal discovery works can be divided into two highly correlated categories i.e., multivariate time series causal discovery, and event sequence causal discovery. However, most previous surveys are only focused on the multivariate time series causal discovery but ignore the second category. In this article, we specify the similarity between the two categories and provide an overview of existing solutions. Furthermore, we provide public datasets, evaluation metrics, and new perspectives for temporal data causal discovery.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Causal reasoning and diagnostics**; • **Mathematics of computing** → **Causal networks**; **Time series analysis;**

Additional Key Words and Phrases: Causal discovery, temporal data analysis, relational learning

Chang Gong and Chuzhe Zhang Equal contributions.

This work has been supported by the **National Natural Science Foundation of China** (**NSFC**) under Grant No. 62472405. And this work has been partially supported by Strategic Priority Research Programs of the Chinese Academy of Sciences under Grant No. XDB0680301 and No. XDB0680101.

Authors' Contact Information: Chang Gong, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China and University of the Chinese Academy of Sciences, Beijing, China; e-mail: gongchang21z@ict.ac.cn; Chuzhe Zhang, School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore; e-mail: felixzh01@outlook.com; Di Yao (Corresponding authors), Institute of Computing Technology Chinese Academy of Sciences, Beijing, China; e-mail: yaodi@ict.ac.cn; Jingping Bi (Corresponding authors), Institute of Computing Technology Chinese Academy of Sciences, Beijing, China; e-mail: bjp@ict.ac.cn; Wenbin Li, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China and University of the Chinese Academy of Sciences, Beijing, China; e-mail: liwenbin20z@ict.ac.cn; YongJun Xu, Institute of Computing Technology Chinese Academy of Sciences, Beijing, China; e-mail: xyj@ict.ac.cn.

ACM 0360-0300/2024/12-ART100
https://doi.org/10.1145/3705297

ACM Reference Format:
Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and YongJun Xu. 2024. Causal Discovery from Temporal Data: An Overview and New Perspectives. *ACM Comput. Surv.* 57, 4, Article 100 (December 2024), 38 pages. https://doi.org/10.1145/3705297

1 Introduction

Temporal data recording the status changing of complex systems is widely collected by different application domains, such as social networks, bioinformatics, neuroscience, finance and climatology, and so on. As one of the most popular data structural, temporal data consists of attribute sequences ordered by time. Owing to the rapid development of sensors and computing devices, research works on temporal data analysis have emerged in recent years. Different approaches have been proposed for different tasks such as classification [90, 197], clustering [196, 199, 201], prediction [163, 164, 200, 204], causal discovery [7, 49, 166], and so on.

Among these tasks, causal discovery detecting the causal relations between many temporal components has become a challenging yet critical task for temporal data analysis. The learned causal structures could be beneficial for explaining the data generation process and guiding the design of data analysis methods. According to the nature and structure of data, the temporal data for causal discovery can be categorized into two groups, i.e., **multivariate time-series (MTS)** and **event sequences (ES)**. MTS data is commonly collected to describe the changes of multiple micro components in various domains, such as finance, weather forecasting, stock market analysis, and industrial process monitoring. By understanding the causal relationships of MTS, we can gain insights into the mechanisms of the system under investigation and guide the design of explainable and robust data analysis models. In contrast, event sequences record the occurrence of multiple events that are organized in a specific order based on their timestamps. Different from MTS, event sequences usually reflect the macro states of a system and the time intervals between events are irregular. By identifying causal relations of events or event types in the multivariate case, we can figure out how events or event sequences influence each other and how they contribute to the future. Thus, casual discovery from temporal data enables us to make better predictions and decisions based on a more accurate causal structure of the systems.

Despite the data, the definitions of causal relations are also diverse. The “Four Causes” theory proposed by Aristotle in philosophy provides an explanation of causality in the world. According to Aristotle, there are four distinct causes, i.e., material cause, formal cause, efficient cause and final cause, that contribute to the existence and understanding of an object or event. Contemporary causal inference methods [141, 145] build upon these philosophical foundations to discover causal relationships, which go beyond correlation and statistical associations to identify the mechanisms and factors that contribute to observations. For different definitions of causality, researchers have developed various techniques for causal discovery. In this article, we focus on the temporal data and summarize four kinds of casual discovery methods suitable for both MTS and ES.

- *Constraint-based Methods.* Causality, in constraint-based methods, refers to the relationship between variables where one variable (the cause) is responsible for the occurrence or the change in another variable (the effect). By testing for conditional independencies in the data, these methods infer the causal structure under the assumption of Pearl’s framework [141]. For example, when two variables are conditionally independent given a set of other variables, it indicates no direct causal relationship between them.
- *Score-based Methods.* Score-based methods aim at finding the graph $\mathcal{G}$ that maximizes the likelihood of the observed data, or its posterior probability given the data, according to the factorization imposed by the graph structure. Instead of the conditional independencies test,

Table 1. Comparison between Existing Surveys and our Work, where CB, SB, GC, SCM, DL, ED, ER, and PA Indicate Constraint-based, Score-based, Granger Causality-based, SCM-based, Deep Learning-based Methods, Evaluation Datasets, Evaluation Results, and Practical Application, Respectively

Survey	Data Type	Multivariate Time-series					Event Sequence				Covered Contents		
		CB	SB	GC	SCM	DL	CB	SB	GC	Others	ED	ER	PA
[78]	Non-temporal	∔	∔	–	∔	–	–	–	–	–	–	✓	–
[61]	Non-temporal	∔	∔	✓	∔	–	–	–	–	–	–	–	✓
[71]	Non-temporal	∔	∔	–	∔	∔	–	–	–	–	✓	–	–
[187]	Non-temporal	∔	∔	✓	∔	∔	–	–	–	–	✓	–	–
[207]	Non-temporal	∔	∔	–	∔	∔	–	–	–	–	✓	–	✓
[103]	Non-temporal	∔	∔	–	–	–	–	–	–	–	✓	–	–
[172]	Non-temporal	∔	∔	–	–	∔	–	–	–	–	–	–	–
[134]	Temporal	✓	–	✓	✓	✓	–	–	–	–	✓	–	–
[166]	Temporal	–	–	✓	–	✓	–	–	✓	–	–	–	–
[138]	Temporal	✓	∔	✓	∔	–	–	–	–	–	✓	–	–
[7]	Temporal	✓	✓	✓	✓	✓	–	–	–	✓	✓	✓	∔
[75]	Temporal	✓	✓	✓	✓	✓	–	–	–	–	✓	✓	✓
Ours	Temporal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	–	✓

"✓", "∔", and "–" indicate that the content is mentioned or briefly mentioned or not mentioned, respectively.

they search causal model with the best score to fit the data as the result. These approaches are suited for scenarios in the presence of complex causal structures and large datasets [61].

- ***Structural Causal Model (SCM)****-based Methods*. In SCM-based methods, causality is defined by a set of structural equations that systematically describe how the value of one variable is generated by the values of other variables [145]. The goal is to construct a model that can not only describe the observed data but also predict the outcomes of potential interventions. This involves both estimating the structure of the causal graph (which variables are direct causes of which other variables) and the parameters of the structural equations (how strong the causal effects are).
- *Granger causality-based Methods*. The Granger causality tests whether one time-series is useful in forecasting another [67]. Granger causality is naturally defined temporarily, i.e., whether the past values of a variable offer additional predictive information beyond its own past values for another variable. It is generally accepted that Granger causality does not capture all aspects of causality but enough to be worth considering for empirical test [169].

This article serves as a systematic and thorough extension of our tutorial [63]. In this article, we focus on two types of temporal data, i.e., MTS data and ES data, and summarize the existing causal discovery methods according to the aforementioned taxonomy. Although the methods of each data cover all four method categories, the research focus varies depending on the type of data. For MTS data, existing researches are balanced in different categories. The constraint-based methods are developed earlier than the other three categories. Thus, there are plentiful constraint-based methods for MTS data. As for ES data, Granger causality-based methods are well-developed since there is a natural match-up between Granger causality and Hawkes processes for ES data. Existing works of both constraint and score-based methods are relatively few but gained increasing attention.

Recently, the progress of causal discovery has been summarized in literature [7, 61, 71, 78, 103, 124, 134, 159, 166, 187]. We compare the representative reviews in Table 1. As shown, these reviews fall into two lines. Reviews in the first line [61, 71, 78, 103, 172, 187, 207] discuss the general causal discovery problem from different perspectives. In these articles, temporal data is taken as one particular application, and many data-specified methods are not included. Reviews in the second line [7, 75, 134, 138, 166] focus on temporal data causal discovery. Methods for bivariate time-series are surveyed in [49]. A thorough overview and recent advances of Granger causality are given in [166]. The recent work [7] establishes the comprehensive taxonomy (i.e., GC, CB, SCM, and SB) for time-series causal discovery, and provides empirical insights. However, very few of these reviews have systematically analyzed methods for event sequence data. To fill the gap, presenting current available methods for both multivariate time-series, and event sequence, we review causal discovery methods from temporal data with the widely accepted taxonomy from [7]. Instead of a complete comparison between the methods, we want to introduce the readers to the methods available in both domains. Specifically, we begin by delineating two tasks in general temporal data through the lens of data types and time interval settings. We then review the currently available methods in both domains, showing that the widely accepted taxonomy from [7], initially applied to time-series, also proves robust for event sequences. Additionally, we discuss the limitations and challenges that vary across different domains.

Nevertheless, causal discovery methods for event sequences are ignored in these reviews. In this article, we not only provide a thoughtful overview of causal discovery methods of the two kinds of temporal data but also give an analysis of the connections and differences between them.

In the following of this survey, we first introduce the background and preliminary of the causal discovery problem in Section 2. The recent progress of causal discovery from multivariate time-series and event sequences are specified in Sections 3 and 4 respectively. After that, we provide resources for performance evaluations in Section 5. Applications, open issues, and new perspectives are provided in Section 6. The whole framework of this survey is shown in Figure 1.

2 Background and Preliminaries

This section begins with the definition of key concepts and assumptions in causal discovery, followed by an overview of three causal graph representations applicable to temporal data. Finally, the problem definitions for causal discovery from MTS and event sequences will be presented.

2.1 Key Concepts and Assumptions in Causal Discovery

Several key concepts serve as the foundation for inferring causal relationships from temporal data. We establish this common ground before discussing research works. Afterward, we present formal definitions for related concepts.

Graph Terminology. A directed **graph** $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ is composed of **nodes V** and edges **edges** $\mathcal{E} \subseteq \mathbf{V}^2$. A node x_i is defined as a **parent** of node x_j if $(x_i, x_j) \in \mathcal{E}, (x_i, x_j) \notin \mathcal{E}$. $\mathcal{G}$ is called a **directed acyclic graph (DAG)** if all edges are directed and there is no directed cycle. The **skeleton** of $\mathcal{G}$ is the version that ignores the directionality of the edges. A **path** in $\mathcal{G}$ denotes a sequence of distinct nodes $x_1, \ldots, x_m$, such that there exits an edge between x_i and x_{i+1} for all $i = 1, \ldots, m-1$. For instance, (x_1, x_2, x_3) consists of a path in Figure 2(c)–(f). Three nodes in $\mathcal{G}$ are called an **immorality** or a **v-structure** in the case that a child of the two others are not adjacent. For example, three nodes in Figure 2(c) consist of a v-structure.

d-separation. In a DAG $\mathcal{G}$, a path between x_1 and x_m is said to be **blocked by a set** S if it satisfies one of the following conditions [141, 145]:

(1) If $x_i \in$ S, the path is blocked if it includes any of these configurations: $(x_{i-1} \rightarrow x_i \rightarrow x_{i+1})$, $(x_{i-1} \leftarrow x_i \leftarrow x_{i+1})$, or $(x_{i-1} \leftarrow x_i \rightarrow x_{i+1})$.
(2) If neither x_i nor any of its descendants is in S, the path is blocked if it forms a v-structure: $(x_{i-1} \rightarrow x_i \leftarrow x_{i+1})$

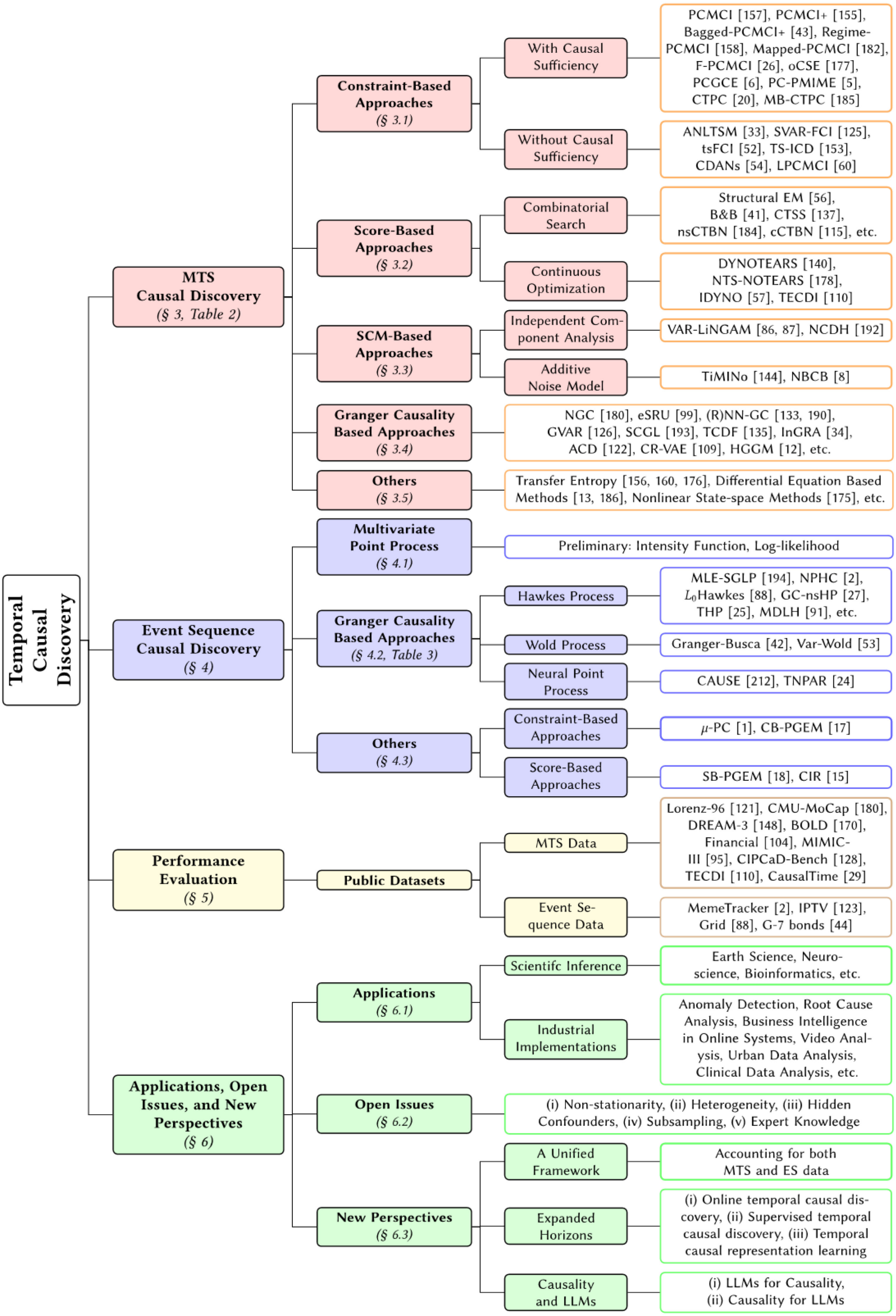
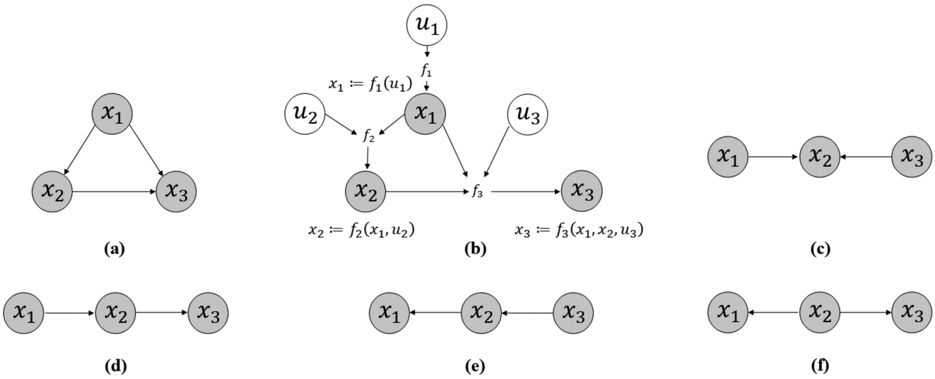

Temporal Causal Discovery
MTS Causal Discovery (§ 3, Table 2)
Constraint-Based Approaches (§ 3.1)
With Causal Sufficiency
PCMCI [157], PCMCI+ [155], Bagged-PCMCI+ [43], Regime-PCMCI [158], Mapped-PCMCI [182], F-PCMCI [26], oCSE [177], PCGCE [6], PC-PMIME [5], CTPC [20], MB-CTPC [185]
Without Causal Sufficiency
ANLTSM [33], SVAR-FCI [125], tsFCI [52], TS-ICD [153], CDANs [54], LPCMCI [60]
Score-Based Approaches (§ 3.2)
Combinatorial Search
Structural EM [56], B&B [41], CTSS [137], nsCTBN [184], cCTBN [115], etc.
Continuous Optimization
DYNOTEARS [140], NTS-NOTEARS [178], IDYNO [57], TECDI [110]
SCM-Based Approaches (§ 3.3)
Independent Component Analysis
VAR-LiNGAM [86, 87], NCDH [192]
Additive Noise Model
TiMINo [144], NBCB [8]
Granger Causality Based Approaches (§ 3.4)
NGC [180], eSRU [99], (R)NN-GC [133, 190], GVAR [126], SCGL [193], TCDF [135], InGRA [34], ACD [122], CR-VAE [109], HGGM [12], etc.
Others (§ 3.5)
Transfer Entropy [156, 160, 176], Differential Equation Based Methods [13, 186], Nonlinear State-space Methods [175], etc.
Event Sequence Causal Discovery (§ 4)
Multivariate Point Process (§ 4.1)
Preliminary: Intensity Function, Log-likelihood
Granger Causality Based Approaches (§ 4.2, Table 3)
Hawkes Process
MLE-SGLP [194], NPHC [2], L_0Hawkes [88], GC-nsHP [27], THP [25], MDLH [91], etc.
Wold Process
Granger-Busca [42], Var-Wold [53]
Neural Point Process
CAUSE [212], TNPAR [24]
Others (§ 4.3)
Constraint-Based Approaches
μ-PC [1], CB-PGEM [17]
Score-Based Approaches
SB-PGEM [18], CIR [15]
Performance Evaluation (§ 5)
Public Datasets
MTS Data
Lorenz-96 [121], CMU-MoCap [180], DREAM-3 [148], BOLD [170], Financial [104], MIMIC-III [95], CIPCaD-Bench [128], TECDI [110], CausalTime [29]
Event Sequence Data
MemeTracker [2], IPTV [123], Grid [88], G-7 bonds [44]
Applications, Open Issues, and New Perspectives (§ 6)
Applications (§ 6.1)
Scientifc Inference
Earth Science, Neuroscience, Bioinformatics, etc.
Industrial Implementations
Anomaly Detection, Root Cause Analysis, Business Intelligence in Online Systems, Video Analysis, Urban Data Analysis, Clinical Data Analysis, etc.
Open Issues (§ 6.2)
(i) Non-stationarity, (ii) Heterogeneity, (iii) Hidden Confounders, (iv) Subsampling, (v) Expert Knowledge
New Perspectives (§ 6.3)
A Unified Framework
Accounting for both MTS and ES data
Expanded Horizons
(i) Online temporal causal discovery, (ii) Supervised temporal causal discovery, (iii) Temporal causal representation learning
Causality and LLMs
(i) LLMs for Causality, (ii) Causality for LLMs

Fig. 1. Framework for causal discovery from temporal data.



(a) (b) (c)

(d) (e) (f)

Fig. 2. Basic DAGs and a simple structural causal model. (a) depicts a DAG structure. (b) illustrates the corresponding SCM of the DAG shown in (a), where x_i and u_i represent endogenous and exogenous variables, respectively, and f_i is the function that determines the values influenced by the structural parents $Pa(x_i)$ and the exogenous variable u_i. (c) illustrates a v-structure. (d)-(f) demonstrate Markov equivalence of graphs.

The concept of blocking in d-separation is crucial for understanding how information is transmitted or obstructed within a DAG. Specifically, if a path between two nodes is blocked by a set S, then S effectively shields x_1 from influencing x_m through that particular path. For instance, in Figure 2(d)–(f), nodes x_1 and x_3 are blocked by S if $x_2 \in$ S. Conversely, in Figure 2(c), x_1 and x_3 are blocked if $x_2 \notin$ S. If two subsets of nodes **A** and **B** are d-separated by a third subset S, it's expressed as

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathrm{S}. \tag{1}$$

Under causal assumptions introduced later in the section, d-separation implies that no causal effect can be transmitted between two sets of nodes if their paths are blocked, making it a valuable tool for causal discovery.

Markov Property. The Markov property is a fundamental assumption in the context of graphical models [106] and causal theory [141, 145]. It relates the notion of graph separation to conditional independencies. This property essentially posits that a joint distribution $P(\mathbf{x})$ over a set of variables $\mathbf{x}$ satisfies the **Markov property** with respect to a DAG $\mathcal{G}$ if it can be factorized according to the structure of the graph. Formally, this is expressed as

$$P(\mathbf{x}) = \prod_{i=1}^{d} P(x_i | Pa(x_i)), \tag{2}$$

where $Pa(x_i)$ denotes the set of parent nodes of x_i in the DAG $\mathcal{G}$, and d is the number of variables. This factorization implies that each variable x_i is conditionally independent of its non-descendants given its parents. Based on the notation ($\perp\!\!\!\perp_{\mathcal{G}}$) in d-separation, we have the equivalence form of Markov property [108] as follows:

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathbf{S} \Rightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{S}. \tag{3}$$

Markov Equivalence of Graphs. Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if and only if they have the same skeleton and share common v-structures [145]. The importance of Markov equivalence in causal discovery lies in its implications for identifying causal relationships from observational (non-experimental) data. It highlights the limitation under some circumstances that not all edge directions in a causal graph can be determined uniquely due to the presence of Markov equivalent graphs.

Markov Blanket and ***Markov Boundary***. Given a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ and a target node Y, the smallest set $\mathbf{MBl}(Y)$ is the **Markov blanket** of Y such that

$$Y \perp\!\!\!\perp \mathbf{V} \backslash (\{Y\} \cup \mathbf{MBl}(Y)) | \mathbf{MBl}(Y). \tag{4}$$

For DAGs, the Markov blanket contains targets' parents, children, and parents of children [141]. What's more, if no proper subset of $\mathbf{MBl}(Y)$ satisfies the property of a Markov blanket of Y, then $\mathbf{MBl}(Y)$ is said to be the Markov boundary [141] of Y. The Markov blanket and Markov boundary hold crucial information to predict the target node, blocking redundant paths. By isolating variables within it, researchers uncover direct influences while reducing the impact of irrelevant nodes.

Causal Sufficiency. If all common causes of all variables are observed, a set of variables is said to be causally sufficient [171], i.e., no hidden confounders. Under this assumption, the majority of causal discovery algorithms presume that the causal structure can be depicted as a DAG.

Faithfulness. For all disjoint node sets $\mathbf{A}, \mathbf{B}, \mathbf{S}$, the independence in probability distribution is faithful [145] to the DAG $\mathcal{G}$ if the following equation holds:

$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{S} \Rightarrow \mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} | \mathbf{S}, \tag{5}$$

where the symbol $\perp\!\!\!\perp_{\mathcal{G}}$ denotes d-separation. While faithfulness is untestable in practice, it is crucial for deriving valid causal inferences from data. If this assumption is violated, the causal relationships become uncertain, posing significant challenges for causal discovery methods [171].

Structural Causal Model (SCM). Pearl's comprehensive theory of causality, as presented in [141], allows to draw causal conclusions from observations using causal hierarchy (PCH) [142]. From that, the structural causal model is defined as a graphical representation of causal relationships that captures how interventions on one or more variables affect the values of other variables in the data generation mechanism. Formally, SCM can be represented in a 4-tuple $< V, U, F, P(U) >$, where V, U denote the set of endogenous and exogenous variables respectively, $P(U)$ is the distribution of exogenous variables, and F represents the set of the mapping function. Specifically, for $f_i \in F$, the model $x_i := f_i(Pa(x_i), u_i), i = 1, \ldots, d$ indicates the assignment of the value x_i to a function of its structural parents $Pa(x_i)$ and exogenous variable u_i. And d denotes the number of variables. For each SCM, we can yield a DAG $\mathcal{G}$ by adding one node for each x_i and directing edges from each parent variable in $Pa(x_i)$ (the causes) to child x_i (the effect). The relationship of the SCM and the corresponding DAG is shown in Figure 2(a) and (b).

Temporal Priority. Temporal priority means that the cause must have occurred before its effect [50]. It is a fundamental assumption of causal discovery from temporal data and creates an asymmetric time relationship in causal processes. However, the time difference between events associated with the time-series may be indistinguishable in cases that the frequencies of data sampling are low. To be specific, dependencies between two variables could be perceived as instantaneous in the observation, leading to *contemporaneous causal relationships* between causes and effects, e.g., $x_2^t \rightarrow x_1^t$ in Figure 4(b).

2.2 Configurations and Causal Structures for Temporal Data

In this section, we begin by outlining two fundamental configurations of temporal data that are crucial for determining the appropriate causal structure and defining the problem. Following this, we present the corresponding causal structures tailored to these temporal data configurations.

Temporal data, based on the collection modes and intrinsic characteristics, can exist in various configurations. These configurations can be categorized as follows:

- **Configuration of Variable Values and State Spaces.** The data collected can be either **discrete** or **continuous**. Discrete values can further be classified into **nominal**, which are

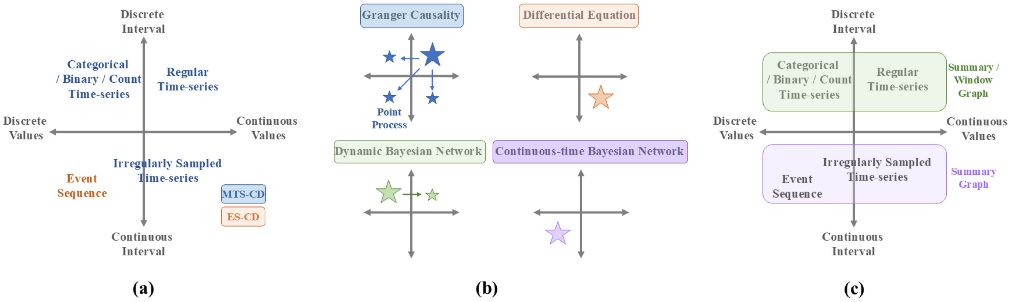


Fig. 3. The configurations for temporal data: (a) illustrates the configuration of time interval and variable values on a coordinate, (b) shows the relation between extant temporal models and temporal data under diverse configurations, (c) depicts the match-up between causal structure and different time-interval configurations.

categories without a specific order, and **ordinal**, which have a defined order [68]. For example, in a healthcare setting, patient symptoms recorded as “mild,” “moderate,” or “severe” represent ordinal data. Continuous data, on the other hand, can assume any value within a numerical range, such as temperature readings [12].

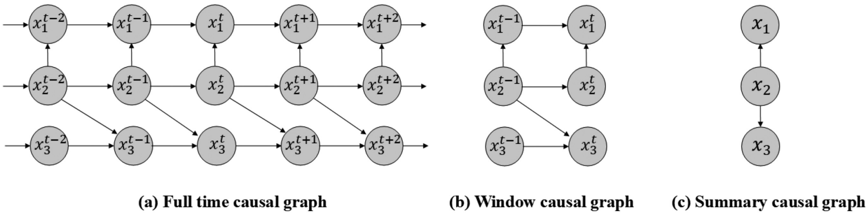
- **Configuration of Time Intervals.** Time intervals can be **continuous** or **discrete**. Continuous intervals imply that data can be recorded at any point in time, offering more fluid and fine-grained data collection approach [117]. Discrete time intervals refer to data being recorded at regular, equally spaced intervals, such as daily or weekly observations. For instance, stock prices recorded at the end of each trading day illustrate discrete time intervals.

As illustrated in Figure 3(a), for ES data, the configuration typically involves a discrete state space coupled with a continuous time interval. In contrast, MTS data can exhibit variable values that are either discrete or continuous. Furthermore, the configuration of time intervals in MTS data is non-trivial to establish. In many scenarios, discrete time intervals are commonly encountered. However, there are also cases where the data is irregularly sampled, resulting in a continuous time interval [117]. This variability in sampling and interval settings plays a significant role in selecting appropriate causal structures and propose proper problem definitions.

For temporal data with diverse configurations, there exist several temporal models such as **Dynamic Bayesian Networks (DBN)** [56, 106, 151], **Continuous-Time Bayesian Networks (CTBN)** [137, 184], Granger causality [67], differential equations [186], the Hawkes process [194], and causal survival analysis [79], which are illustrated in Figure 3(b), each providing causal explanations or interpretations under specific assumptions. These models offer different perspectives on capturing temporal dependencies and causal relationships in data.

To align the results from these various temporal models for the purpose of causal structure learning, and to provide a systematic review, we follow [6, 7, 145] and provide a unified description of a few key causal structures applicable to temporal data, i.e., *full-time causal graph*, *window causal graph*, and *summary causal graph*.

Full-time Causal Graph. As illustrated in Figure 4(a), the *full-time causal graph* [7] represents a complete graph of the dynamic systems. It can also be understood as analogous to an unrolled network of Dynamic Bayesian Networks [106], which visualizes the repeated structure of a dynamic model across multiple time slices. For d-variate time-series $\mathbf{x}$, the measurement at each time point t is a vector $(x_1^t, \dots, x_d^t)$. Vertices in full-time causal graphs consist of the set of component $x_1, \dots, x_d$ at each time point t with lag-specific directed links such as $x_i^{t-k} \longrightarrow x_j^t$. However, it is usually difficult to discover full-time causal graphs due to the single observation for each series at each time point.



(a) Full time causal graph **(b) Window causal graph** **(c) Summary causal graph**

Fig. 4. Causal graphs for discrete-time observations.

Window Causal Graph. To remedy this problem, by assuming a time-homogeneous causal structure, *window causal graph* [6, 7, 145] is proposed. To be specific, the dynamics of observation $\mathbf{x}$ are governed by $\mathbf{x}^t := f(\mathbf{x}^{<t}, \mathbf{u}^t)$ where the function f indicates the future observation based on past $\mathbf{x}^{<t}$ and the exogenous term $\mathbf{u}^t$. As illustrated in Figure 4(b), the window causal graph is represented in a time window, the size of which amounts to the maximum lag in the full-time causal graph. Due to the constraints imposed by discrete time interval settings, the window causal graph is specifically suited for temporal data with discrete time intervals, as illustrated in Figure 3(c).

Summary Causal Graph. As shown in Figure 4(c), each node from the *summary causal graph* is in the collapsed form from time-series. That is to say, causal relations between time-series are illustrated in the summary graph without referring to time lags [144]. In many applications, it is adequate to model the relationships between temporal variables without having precise knowledge of the interactions at specific time instants. The summary causal graph, which provides a high-level representation of causal relations, is applicable to temporal data with all configurations mentioned above (i.e., continuous and discrete time intervals). This broad applicability makes the summary causal graph an ideal result for causal discovery and a versatile tool for downstream analyses.

2.3 Problem Definitions

Based on the taxonomy and causal structures for temporal data outlined above, the task of causal discovery from temporal data can be categorized into two distinct problems, as depicted in Figure 5. These are **causal discovery from multivariate time-series (MTS-CD)** and **causal discovery from event sequences (ES-CD)**.

Definition 2.1 (MTS-CD). Consider a time-series with d variables: $\{\mathbf{x}^t\}_t = \{(x_1^t\ x_2^t\ \dots\ x_d^t)^\top\}_t$. We assume that causality between variables are entailed by the following structural causal model:

$$x_i^t := f_i(Pa(x_i^t), u_i^t),\ i = 1, \dots, d, \tag{6}$$

where for any $i \in \{1, \dots, d\}$ at time instance t, $Pa(x_i^t)$ is the set of direct parents of x_i^t which can be both in the past and at the same time instance. u_i^t denotes the independent noise [143]. Without loss of generality, for $\{\mathbf{x}^t\}_{t\in\mathbb{Z}^+}$, where $\mathbb{Z}^+$ indicates the set of all positive integers, discrete time interval MTS data is described. In contrast, for $\{\mathbf{x}^t\}_{t\in\mathbb{R}^+}$, where $\mathbb{R}^+$ indicates the set of all positive real numbers, this definition is appropriate for continuous time interval configurations. MTS-CD aims at finding either of the two kinds of outputs, i.e., summary causal graph (A) or window causal graph (W and A^k).

As for the summary causal graph, the output is the adjacency matrix $A \in \mathbb{Z}^{d\times d}$ which summarizes the causal structure, and the (i, j)-th entry of the matrix A is 1 if past observations of x_i enter the structural equation of x_j^t and 0 otherwise. We say that 'x_i causes x_j' if $A_{ij} = 1$. As for the window causal graph with a maximum time lag K, the output matrices W and A^k ($k \in \{1, \dots, K\}$) correspond to intra-slice and inter-slice edges, respectively. For example, $W_{ij} = 1$ denotes the

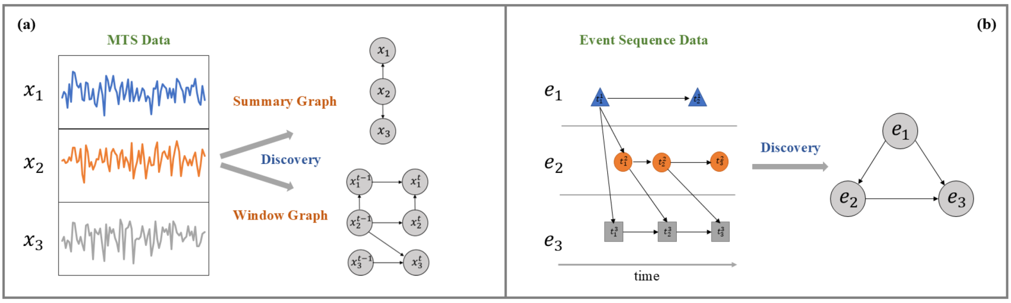


Fig. 5. The problem description: (a) causal discovery from MTS, (b) causal discovery from event sequence.

instantaneous dependency $x_i^t \to x_j^t$, while $A_{ij}^k = 1$ denotes a lagged dependency $x_i^{t-k} \to x_j^t$ for $k > 0$.

Definition 2.2 (ES-CD). An event sequence is denoted by $\{(t_n, e_n)\}_{n=1}^N$ with $t_1 < \ldots < t_N$, where t_n is the timestamp of the nth event and e_n stands for the corresponding event type. ES-CD aims at deriving a causal graph parameterized by an adjacency matrix $A \in \mathbb{Z}^{E \times E}$ where E is the number of event types. If $A_{ij} = 1$, we say event-type i is a cause of event-type j.

3 Causal Discovery from Multivariate Time-series

As stated, existing MTS-CD methods can be divided into five categories, including constraint-based, score-based, SCM-based, GC-based, and other methods. Given time-series data, MTS-CD involves identifying different causal graphs (i.e., *summary causal graph* or *window causal graph*) under complex situations (e.g., *non-linearity*, *hidden confounders*). Table 2 utilizes the taxonomy from [7] and incorporates emerging new methods and additional perspectives to comprehensively summarize the details of all MTS-CD methods reviewed in this section.

3.1 Constraint-based Methods

Constraint-based MTS-CD methods depend on statistical tests of conditional independence. The general steps encompass the following: (i) *graph skeleton discovery*, (ii) *rule-based orientation*. Firstly, they establish a skeletal structure among variables through the application of conditional independence. Secondly, the orientation of the skeleton is determined based on rules, for example, cycles or new v-structures should be avoided. Most constraint-based MTS-CD methods can construct the window causal graph. And they assume causal Markov property and faithfulness. However, they exhibit variations in their requirements regarding the assumption of causal sufficiency (i.e., no hidden confounders). As the presence of hidden confounders can significantly impact the validity and reliability of results, we will review methods with and without causal sufficiency, separately.

3.1.1 Methods with Causal Sufficiency. Most methods in this subclass are temporal variants of the ***Peter-Clark (PC)*** algorithm [171], which assumes causal sufficiency. Endeavors [6, 20, 155, 157] for MTS data extend *PC*'s ability to model temporal context by considering autocorrelation, non-stationarity, and high dimensionality.

(1) ***PCMCI** and its Variants*. Runge et al. [157] propose *PCMCI*, which starts by constructing a partially connected graph, where all pairs of nodes (x_i^{t-k}, x_j^t) are directed as $x_i^{t-k} \to x_j^t$ accounting for temporal priority. The algorithm comprises two stages: (i) *PCMCI* eliminates redundant edges based on conditional independence and further removes edges by assuming

Table 2. Summary of MTS-CD Methods Reviewed in this Survey

Section	Method	Target		Assumption			Configuration		Issues					
		Graph	Contemporaneous	Sufficiency	Markov	Faithfulness	Value	Interval	Nonlinear	Hidden Conf	Intervention	Heterogeneity	Non-stationarity	Irregular Sampling
CB	PCMCI (2019) [157]	Window	No	Yes	Yes	Yes	Both	Discrete	Yes	No	No	No	No	No
	PCMCI$^+$ (2020) [155]	Window	Yes	Yes	Yes	Yes	Both	Discrete	Yes	No	No	No	No	No
	Bagged-PCMCI$^+$ (2023) [43]	Window	Yes	Yes	Yes	Yes	Both	Discrete	Yes	No	No	No	No	No
	Regime-PCMCI (2020) [158]	Window	No	Yes	Yes	Yes	Both	Discrete	Yes	No	No	No	Yes	No
	F-PCMCI (2023) [26]	Window	No	Yes	Yes	Yes	Both	Discrete	Yes	No	No	No	No	No
	oCSE (2015) [177]	Summary	No	Yes	Yes	Yes	Continuous	Discrete	Yes	No	No	No	No	No
	PCGCE (2022) [6]	Extended	Yes	Yes	Yes	Yes	Continuous	Discrete	Yes	No	No	No	No	No
	PC-PMIME (2023) [5]	Summary	No	Yes	Yes	Yes	Continuous	Discrete	Yes	No	No	No	No	No
	CTPC (2021) [20]	Summary	No	Yes	Yes	Yes	Discrete	Continuous	Yes	No	No	No	No	Yes
	MB-CTPC (2023) [185]	Summary	No	Yes	Yes	Yes	Discrete	Continuous	Yes	No	No	No	No	Yes
	ANLTSM (2008) [33]	Window	Yes	No	Yes	Yes	Continuous	Discrete	Yes	Yes	No	No	No	No
	tsFCI (2010) [52]	Window	No	No	Yes	Yes	Both	Discrete	Yes	Yes	No	No	No	No
	SVAR-FCI (2018) [125]	Window	Yes	No	Yes	Yes	Continuous	Discrete	No	Yes	No	No	No	No
	TS-ICD (2023) [153]	Window	Yes	No	Yes	Yes	Both	Discrete	No	Yes	No	No	No	No
	CDANs (2023) [54]	Window	Yes	No	Yes	Yes	Continuous	Discrete	Yes	Yes	No	No	Yes	No
	LPCMCI (2020) [60]	Window	Yes	No	Yes	Yes	Both	Discrete	Yes	Yes	No	No	No	No
SB	DYNOTEARS (2020) [140]	Window	Yes	Yes	Yes	No	Both	Discrete	No	No	No	No	No	No
	NTS-NOTEARS (2021) [178]	Window	Yes	Yes	Yes	No	Both	Discrete	Yes	No	No	No	No	No
	IDYNO (2022) [57]	Window	Yes	Yes	Yes	No	Continuous	Discrete	Yes	No	Yes	No	No	No
	TECDI (2023) [110]	Window	Yes	Yes	Yes	No	Continuous	Discrete	Yes	No	Yes	No	No	No
SCM	VAR-LiNGAM (2008) [86]	Window	Yes	Yes	Yes	No	Continuous	Discrete	No	No	No	No	No	No
	NCDH (2022) [192]	Summary	No	Yes	Yes	No	Continuous	Discrete	Yes	No	No	No	Yes	No
	TiMINo (2013) [144]	Summary	Yes	Yes	Yes	Yes	Continuous	Discrete	Yes	No	No	No	No	No
	NBCB (2021) [8]	Summary	Yes	Yes	Yes	Yes	Continuous	Discrete	Yes	No	No	No	No	No
GC	(R)NN-GC (2015,2018) [133, 190]	Summary	Yes	No	No	No	Continuous	Discrete	Yes	No	No	No	No	No
	NGC (2022) [180]	Summary	No	No	No	No	Continuous	Discrete	Yes	No	No	No	No	No
	eSRU (2020) [99]	Summary	No	No	No	No	Continuous	Discrete	Yes	No	No	No	No	No
	SCGL (2019) [193]	Summary	No	No	No	No	Continuous	Discrete	Yes	No	No	No	No	No
	GVAR (2021) [126]	Summary	No	No	No	No	Continuous	Discrete	Yes	No	No	No	No	No
	TCDF (2019) [135]	Window	Yes	No	No	No	Continuous	Discrete	Yes	Yes	No	No	No	No
	CR-VAE (2023) [109]	Summary	Yes	No	No	No	Continuous	Discrete	Yes	No	No	No	Yes	No
	InGRA (2020) [34]	Summary	No	No	No	No	Continuous	Discrete	Yes	No	No	Yes	No	No
	ACD (2022) [122]	Summary	No	No	No	No	Continuous	Discrete	Yes	Yes	No	Yes	No	No
	HGGM (2019) [82]	Summary	No	No	No	No	Continuous	Discrete	Yes	No	No	Yes	No	No
Others	DBCL (2010) [186]	Summary	Yes	No	Yes	Yes	Continuous	Discrete	Yes	Yes	No	No	No	No
	NGM (2022) [13]	Summary	Yes	No	No	No	Continuous	Continuous	Yes	No	No	No	No	Yes
	CCM (2012) [175]	Summary	No	No	No	No	Continuous	Discrete	Yes	No	No	No	No	No

consistency across time. (ii) The application of **Momentary Conditional Independence (MCI)** [147] addresses auto-correlation and prevents the occurrence of spurious correlations. In this context, MCI serves as a measure that conditions on the parents of x_j^t and x_i^{t-k} while assessing $x_i^{t-k} \not\perp\!\!\!\perp x_j^t | Pa(x_j^t) \backslash \{x_i^{t-k}\}, Pa(x_i^{t-k})$. And $\not\perp\!\!\!\perp$ denotes non-independence.

Recent years have witnessed notable improvement endeavors in *PCMCI*. In ***PCMCI+***, Runge [155] expands *PCMCI*'s capability to model contemporaneous relations. In ***Bagged-PCMCI***+, Debeire et al. [43] enhance the stability of results through bootstrap sampling and assigns confidence to edges. In ***Regime-PCMCI***, Saggioro et al. [158] focus on regime-dependent causal relations, and iteratively optimize regimes and causal graphs in each regime. In ***Mapped-PCMCI***, Tibau et al. [182] extract causal relations from MTS at the grid level. In ***F-PCMCI***, Castri et al. [26] employ a feature selection approach to enhance *PCMCI*'s performance in handling a large number of variables.

(2) ***oCSE** and its Variants*. Sun et al. [177] propose *oCSE* to guide computational and efficient MTS-CD. It builds upon the theoretical framework of **Causation Entropy (CE)** [176], an extension of **Transfer Entropy (TE)** [160] that enables the quantification of network relationships among multiple variables rather than just pairwise relations. Slightly different from *PC*, *oCSE* encompasses the past of all accessible nodes, rather than minimizing the size of the conditioning set by conditioning on all potential causes.

oCSE is a computational and sample-efficient algorithm. However, it assumes that the hidden dynamics follow a stationary first-order Markov process as the CE only models causal relations with time lags equal to one. Recently, Assaad et al. [6] propose ***PCGCE*** to extract extended summary causal graphs based on *PC* and the Greedy Causation Entropy, which is a variant of CE modeling more intricate historical information. In ***PC-PMIME***, Arsac et al. [5] combine *PC* with an information-theoretic measure (i.e., PMIME) designed to detect direct coupling in time-series.

(3) ***CTPC*** and its Variants. Bregoli et al. [20] propose ***Continuous-Time PC (CTPC)***, which is the first constraint-based approach for the structure learning of CTBNs. On the basis of *PC*,

CTPC consists of a proper set of statistics testing conditional independence for time-series data with continuous-time observations. Compared to its score-based counterpart [137], *CTPC* has better accuracy in terms of structure learning when variables in the CTBN have more than two values. Recently, ***Markov Blanket-based Continuous-Time PC (MB-CTPC)*** [185] adapts *CTPC* and only tests dependencies relevant to the Markov blanket of specified variables, which is helpful in reducing computational time in downstream tasks such as multivariate time-series classification.

3.1.2 Methods without Causal Sufficiency. Prior methodologies within this subclass involve temporal adaptations of the ***FCI*** algorithm [171], which does not assume causal sufficiency. However, recent years have witnessed the emergence of alternative methods that do not rely on the *FCI* premise. In this subsection, we will begin by providing a concise overview of *FCI* before exploring its variants for MTS-CD. Subsequently, we will review methods that operate independently of *FCI*.

FCI offers a broader applicability by accommodating hidden confounders. It's demonstrated to be accurate, sound, complete, and consistent from a theoretical standpoint [210]. It accomplishes this property through (i) the partial ancestral graphs for representing the presence of hidden confounders, and (ii) the incorporation of supplementary rules [209] for edge orientations.

The variants of *FCI* extend its ability to model temporal information in MTS.

(1) ***ANLTSM** and its Variants*. Chu and Glymour [33] propose *ANLTSM*, which is a constraint-based additive nonlinear time-series model. It presumes the following data generation process to escape the curse of dimensionality for nonparametric conditional independence tests. The *FCI* algorithm is then leveraged to identify lagged and contemporaneous causal relations. *ANLTSM* imposes constraints on contemporaneous interactions, requiring them to be linear, and on hidden confounders, stipulating them to be linear and contemporaneous. Following that, there are subsequent endeavors, building upon a model-defined data generation process. In ***SVAR-FCI***, Malinsky and Spirtes [125] assume that the data generation process for MTS adheres to a **structural vector autoregressive** (**SVAR**) model and takes into account contemporaneous influences.

(2) ***tsFCI***. Entner and Hoyer [52] propose ***time-series FCI (tsFCI)*** for MTS, which directly applies *FCI* based on a time window. In detail, the initial time-series data is transformed into a collection of samples of the random vector using a sliding window of size K. Subsequently, by treating each element of the transformed vector as an individual random variable, *FCI* can be directly applied. Nevertheless, *tsFCI* overlooks selection variables and contemporaneous causal relationships.

In recent years, several methods have emerged that employ alternative approaches to handle hidden confounders, distinct from *FCI*.

(3) ***TS-ICD***. Rohekar et al. [153] propose ***time-series iterative causal discovery (TS-ICD)*** which serves as an efficient approach for MTS-CD while accounting for hidden confounders. The core concept of *TS-ICD* encompasses two primary elements: (i) Drawing inspiration from [152], *TS-ICD* adopts an iterative refinement strategy for a graph recovered through preceding iterations with smaller conditioning sets, thereby enhancing statistical power and promoting stability in contrast to *FCI*. (ii) In the context of temporal data, *TS-ICD* first learns long-term temporal relations then short term ones, and contemporaneous relations are inferred lastly.

(4) ***CDANs***. In order to address the presence of changing modules (i.e., non-stationarity), Ferdous et al. [54] propose a framework for ***causal discovery from autocorrelated and non-stationary time-series (CDANs)***. Their rationale is grounded on the notion that disregarding the presence of changing modules may give rise to erroneous or misleading causal links in MTS-CD. Specifically, *CDANs* incorporates a proxy variable that captures the variability of changing modules over time. Subsequently, kernel-based conditional independence tests are employed to identify contemporaneous relationships between the proxy variables and the observed variables.

(5) ***LPCMCI***. In order to expand the applicability of *PCMCI* method to situations involving latent confounders, Gerhardus and Runge [60] propose an extension known as *LPCMCI*. It introduces novel concepts, namely middle marks and LPCMCI-PAGs, which provide an unambiguous causal interpretation to facilitate the orientation of edges in the presence of hidden confounders. Furthermore, it mitigates the issue of inflated false positives.

3.2 Score-based Methods

In this section, the main ideas of score-based approaches will first be presented, including model scoring, and model search. Then, we will review combinatorial search approaches and continuous optimization approaches for MTS-CD, respectively.

3.2.1 Main Ideas of Score-based Methods. The intuition of score-based approaches is to find the best graph structure measuring its fitness to data [106, 159]. The graph structures in these approaches are often modeled as Bayesian Networks or their variants [56, 151, 184] for temporal data, which provide causal explanations under causal edge assumptions [141]. It consists of two elements: (i) *model scoring*, and (ii) *model search*.

Model Scoring. Score functions play a crucial role in ranking the inferred graphs based on data. Common objective functions can be categorized into two groups [103], i.e., the Bayesian scores and information-theoretic scores. The Bayesian scores are based on both goodness-of-fit and the incorporation of prior knowledge, including BDe scores, K2 scores, and so on [98]. The information-theoretic scores take the model complexity into account to avoid over-fitting, including BIC and AIC scores [22].

Model Search/Optimization. This process can be additionally divided into two stages, i.e., parent set identification and structure optimization [159]. In the first stage, parent set identification generates a ranked list of suitable candidate parent sets for each variable, with higher-scoring sets listed first. Most approaches concentrate on the second stage, which cast the problem of searching causal structure $\mathcal{G}$ into an optimization problem using the aforementioned score functions S. As stated in [145], the ultimate goal is: $\hat{\mathcal{G}} = \text{argmin}_{\mathcal{G} \text{ over } \mathbf{x}} S(\mathcal{D}, \mathcal{G})$, where the empirical data $\mathcal{D}$ represents variables $\mathbf{x}$. Traditionally, it involves a combinatorial graph search, which is known to be a time-consuming problem. And currently the state-of-the-art on structure optimization is represented by hill climbing and tabu search [37, 162]. Nonetheless, causal discovery algorithms remain vulnerable to the challenges posed by the curse of dimensionality, which can result in suboptimal solutions that fail to fully capture the true underlying causal relationships. Recently, in the context of ***NOTEARS*** [213], an algebraic breakthrough has been utilized to characterize the acyclicity constraint in structure learning. This advancement transforms the previously *discrete combinatoric search problem* into a *continuously optimizing problem*. Consequently, the convertible relationship ($\Longleftrightarrow$) can be formulated as follows:

$$\begin{array}{c} \min_{\mathbf{A}\in\mathbb{R}^{d\times d}} S(\mathbf{A}) \\ \text{subject to } \mathcal{G}(\mathbf{A}) \in \text{DAGs} \end{array} \quad \Longleftrightarrow \quad \begin{array}{c} \min_{\mathbf{A}\in\mathbb{R}^{d\times d}} S(\mathbf{A}) \\ \text{subject to } h(\mathbf{A}) = 0, \end{array} \tag{7}$$

where $\mathbf{A}$ denotes the adjacency matrix, and $h(\cdot)$ function enforces acyclicity in structure learning. The original acyclicity constraint function is implemented as $h(\mathbf{A}) = \text{tr}(e^{\mathbf{A}\circ\mathbf{A}}) - d$, where $\circ$ is the Hadamard product. The equation restricts that there exist no loop within any length of steps based on the Taylor expansion [213]. It relies on the augmented Lagrangian method to solve the continuous constrained optimization problem.

In the context of MTS, we review the score-based methods following a similar paradigm from combinatoric search to continuous optimization.

3.2.2 Combinatorial Search Methods. It's non-trivial to conduct the combinatorial graph search as it's reported to be an NP-hard problem [31, 174]. Researchers have developed various score-based approaches, including those for DBNs [41, 56] and CTBNs [36, 115, 137, 184].

For learning DBNs, Friedman et al. [56] propose ***Structural Expectation-Maximization (Structural-EM)***. EM is concerned with learning BN models when hidden variables are assumed to be present [55]. And *Structural-EM* can learn DBN given even partial observations of variables over time. Yu et al. [205] propose an influence score that advances the capability of DBN to identify the +/− sign, such as activation or repression in biological data, and the relative magnitude between interactive temporal variables. In [151], Robinson and Hartemink introduce a non-stationary dynamic Bayesian network specifically designed for scenarios where the assumption that time-series data is generated by a stationary process is violated. Grzegorczyk and Husmeier [69, 70] propose a series of works learning DBNs under non-stationary scenarios. In ***B&B***, Campos and Ji [41] first cast the problem of structure learning in DBN into a corresponding augmented BN using structural constraints, then use a branch-and-bound approach to guarantee global optimality. The *B&B* algorithm exhibits two notable properties regarding its computational complexity. First, it is an anytime and exact method, whereby the algorithm can be terminated at any point to return the current best solution along with an upper bound on the globally optimal value. Second, the algorithm can be easily parallelized and can provide a linear speedup in the number of tasks.

For learning CTBNs, Nodelman et al. [137] first propose a score-based approach called ***Continuous-Time Search and Score (CTSS)***. *CTSS* defines a Bayesian score for evaluating different graph candidates, and leverages a greedy hill-climbing search to find the structure with a high score. As there are no acyclicity constraints for CTBNs, the heuristic search process is much more efficient than that for DBNs [36, 137]. For non-stationary data, Vila and Stella [184] propose ***non-stationary CTBN (nsCTBN)*** with score functions and the corresponding optimization algorithm covering different non-stationary scenarios in a systematical manner. Recently, Linzner and Koeppl [115] introduce ***conditional CTBN (cCTBN)*** for interventional data in an active learning manner.

3.2.3 Continuous Optimization Methods. The majority of approaches in this subclass are temporal adaptations derived from *NOTEARS*, as discussed in Section 3.2.1. In the context of time-series data, the reviewed methods enhance the abilities of *NOTEARS* by incorporating temporal context.

(1) ***DYNOTEARS***. Pamfil et al. [140] propose *DYNOTEARS*, which captures linear causal relations from MTS based on a continuous optimization approach. A structural vector autoregressive model assumption is introduced, which reflects both contemporaneous and time-lagged causal effects:

$$\mathbf{x}^t = \mathbf{x}^t \mathbf{W} + \mathbf{x}^{t-1} \mathbf{A}^1 + \ldots + \mathbf{x}^{t-K} \mathbf{A}^K + \mathbf{u}^t, \tag{8}$$

where K is the model order, $\mathbf{u}$ is a vector of centered error variables. To guarantee the identifiability, the error terms $\mathbf{u}^t$ are assumed either non-Gaussian or standard Gaussian to hold identifiability [145]. $\mathbf{W}$ and $\mathbf{A}$ are weighted adjacency matrices, which correspond to intra-slice edges (contemporaneous relationships) and inter-slice edges (time-lagged relationships), respectively. The procedure of structure learning revolves around minimizing the least-squares loss subject to an acyclicity constraint, and the optimization problem is defined as follows:

$$\begin{gathered} \min\nolimits_{\mathbf{W},\mathbf{A}} \ f(\mathbf{W},\mathbf{A}) \ \text{ s.t. } h(\mathbf{W}) = 0, \\ \text{where } \ f(\mathbf{W},\mathbf{A}) = \frac{1}{2n} ||\mathbf{X}^t - \mathbf{X}^t \mathbf{W} - \mathbf{X}^{(t-K):(t-1)} \mathbf{A}||_F^2 + \lambda_{\mathbf{W}} ||\mathbf{W}||_1 + \lambda_{\mathbf{A}} ||\mathbf{A}||_1, \end{gathered} \tag{9}$$

where n is the number of timesteps, and $\lambda_{\mathbf{W}}, \lambda_{\mathbf{A}}$ stand for weights of regularization. The acyclicity constraint is imposed on the contemporaneous relationships $\mathbf{W}$. As a result, the window causal graph can be estimated by optimizing $\mathbf{W}$ and $\mathbf{A}$.

(2) ***NTS-NOTEARS.*** To extract both linear and nonlinear relations among variables, Sun et al. [178] propose *NTS-NOTEARS* which leverages 1D convolutional neural networks. The optimization procedure follows a similar way as *DYNOTEARS*. Besides, prior knowledge of variable dependencies can be transformed as additional optimization constraints.

(3) ***IDYNO** and its Variants.* To handle both observational and interventional MTS data, Gao et al. [57] propose an ***interventional extension of DYNOTEARS (IDYNO)***, which modifies the score function to handle interventional targets. In ***Temporal Causal Discovery based on Intervention (TECDI)*** [110], Li et al. propose scores for imperfect intervention.

Nevertheless, it is important to acknowledge the boundaries and limitations of continuous optimization methods discussed in [96, 136, 149], such as the impact of the data scale and the convergence criteria of the augmented Lagrangian method. It is recommended to consider these aspects for future advancements and applications of this methodology.

3.3 Structural Causal Model-based Methods

SCM-based methods leverage the structural causal model, such as $x_j = f_j(x_i, u_j)$, to represent the data. These methods allow for the inference of causal relationships by taking into account the effects of noise distortion u or non-linearity in the causal functions f [35, 66, 145]. When applied to time-series data, SCM-based methods can be broadly categorized into two groups: (i) independent component analysis-based methods and (ii) other methods. In the upcoming sections, we will review these two categories in detail.

3.3.1 Independent Component Analysis-based Methods. The methods in this subclass are temporal adaptations derived from the ***LiNGAM*** algorithm [165]. Prior to delving into the specific methods for MTS-CD, it is essential to provide a concise overview of *LiNGAM*.

Shimizu et al. [165] propose causal discovery based on ***Linear, Non-Gaussian, Acyclic Models (LiNGAM)*** in non-temporal setting, which has the following assumptions: (i) a linear data generation process, (ii) non-Gaussian disturbances, and (iii) no hidden confounders. In *LiNGAM*, the relations among observations can be formulated as $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{u}$, where $\mathbf{A}$ denotes the adjacency matrix of the causal graph. The equation can be rewritten as $\mathbf{x} = \mathbf{B}\mathbf{u}$, where $\mathbf{B} = (\mathbf{I} - \mathbf{A})^{-1}$. Then, *LiNGAM* considers $\mathbf{x}$ as observed signals and $\mathbf{u}$ as underlying independent components. To uncover hidden factors or sources that contribute to the observed signals, **independent component analysis (ICA)** [173] can be leveraged to estimate a linear transformation from $\mathbf{u}$ to $\mathbf{x}$, *i.e.* $\mathbf{B}$. Then causal relationships $\mathbf{A}$ can be derived based on $\mathbf{B}$. MTS-CD methods based on ICA are reviewed. They extend *LiNGAM*'s ability to model temporal context, non-linearity, and time-varying situations.

(1) ***VAR-LiNGAM** and its Variants.* Hyvärinen et al. [86, 87] propose *VAR-LiNGAM* for MTS-CD. It assumes the utilization of a SVAR model as SCM, incorporating the non-Gaussian nature of causal process. It's defined as $\mathbf{x}^t = \sum_{k=0}^{K} \mathbf{A}^k \mathbf{x}^{t-k} + \mathbf{u}^t$, where $\mathbf{A}^k$ is the adjacency matrix of the causality between the variables $\mathbf{x}$ with time lag k. And $\mathbf{u}^t$ are independent random variables modeling the exogenous influences, which are assumed to be non-Gaussian. To derive causal adjacency matrix $\mathbf{A}$, *VAR-LiNGAM* takes a four-step optimization procedure similar to *LiNGAM*, i.e., (i) regression phase, (ii) residual computation, (iii) contemporaneous relation inference based on *LiNGAM*, (iv) delayed relation inference based on matrix transformation. It degenerates to the *LiNGAM* model if the autoregressive order is zero, i.e., $K = 0$. The *VAR-LiNGAM* model is further extended to challenging scenarios, such as time-varying situations [83], cyclic structures [107].

(2) ***NCDH***. Wu et al. [192] propose ***Nonlinear Causal Discovery via HM-NICA (NCDH)***, which leverages a nonlinear ICA algorithm as a measurement of nonlinear relationships. *NCDH* first leverages the combination of nonlinear ICA and hidden Markov model to separate latent noises [72]. As a remedy for the permutation uncertainty of ICA, a sequence of independence tests are conducted to determine the corresponding relations between the observed variables and the separated noises.

3.3.2 Other SCM-based Methods. Except for ICA-based methods, several works in MTS-CD have emerged that take into account the effects of noise distortion or non-linearity in the causal functions.

(1) ***TiMINo** and its Variants*. Peters et al. [144] propose ***time-series Models with Independent Noise (TiMINo)***, which fits SCM for time-series data. *TiMINo* outputs either a summary time graph or remains undecided, which avoids leading to wrong causal conclusions when the model is misspecified or the data is insufficient.

(2) ***NBCB***. Assaad et al. [8] propose ***Noise-Based/Constraint-Based (NBCB)*** approach. It first leverages an SCM-based method to infer potential causal relationships. The data generation process is defined as $x_j^t = f_j(Pa(x_j^{\tau})^{t-K}, \dots, Pa(x_j^0)^t) + u_j^t$, and can be estimated with a Gaussian process. Spurious edges are subsequently eliminated by taking into consideration the set of potential parental relationships. Recently, Bystrova and Assaad et al. [23] propose hybrid approaches, combining SCM-based and constraint-based approaches. They aim at exploiting the best of two groups of endeavors and show robustness to assumption violations.

3.4 Granger Causality-based Methods

In this section, we will commence by presenting the preliminaries of **Granger Causality (GC)**, encompassing its original definition and associated concepts. Granger causality analysis, initially introduced by Granger (1969) in the seminal work [67], represents a powerful approach for discerning causal relationships. It's based on testing the ability of one variable to improve the forecasting of another variable compared to using only the past information of the latter. The detailed definition is as follows:

Definition 3.1 (GC between time-series). A time-series x_i Granger-causes x_j if past values of x_i provide unique, statistically significant information about future values of x_j. According to this proposition, x_i is defined to be "causal" for x_j if

$$\mathrm{var}\left[x_j^t - \mathcal{P}\left(x_j^t | \mathcal{H}^{<t}\right)\right] < \mathrm{var}\left[x_j^t - \mathcal{P}\left(x_j^t | \mathcal{H}^{<t} \backslash x_i^{<t}\right)\right], \tag{10}$$

Here, $\mathcal{P}(x_j^t | \mathcal{H}^{<t})$ represents the optimal prediction of x_j^t considering the historical information $\mathcal{H}^{<t}$, which encompasses all relevant data up until time t. Additionally, $\mathcal{H}^{<t} \backslash x_i^{<t}$ denotes the exclusion of the information pertaining to $x< t$ from the historical information set $\mathcal{H}_{<t}$.

Based on the original definitions, there also exist diverse formulations of Granger causality that incorporate various model specifications and statistical methodologies to enhance representational capacity and facilitate inference, such as autoregressive models [67] and neural networks [180]. Previously, the majority of techniques employed to identify Granger causality in bivariate scenarios have been extensively explored. In the subsequent section, our focus will shift toward reviewing methodologies applicable to multivariate contexts, which can be classified into two main categories: statistical learning-based approaches and deep learning-based approaches.

Subsequently, we will proceed to review statistical learning-based methods and deep learning-based methods for MTS-CD.

3.4.1 Statistical Learning-based Methods. Statistical learning-based methods encompass two main categories: (i) *model-free approaches* and (ii) *model-based approaches*. Model-free approaches involve estimating conditional probability density functions [10] or utilizing transfer entropy [101] to infer Granger causal relationships. These approaches exploit the ability to model nonlinear dependencies, but encounter challenges when the number of variables increases due to the curse of dimensionality. In contrast, model-based methods offer computational efficiency in high-dimensional settings by employing parameterized data generative models, such as VAR models [4] or kernel models [150], to effectively capture the dynamics of the measured time-series.

Due to space limitations, we leave details of statistical learning-based methods, including model-free and model-based methods, in supplementary materials.

3.4.2 Deep Learning-based Methods. Neural networks possess the capability to capture intricate, nonlinear, and non-additive relationships between input and output variables. In order to infer Granger causal relationships within nonlinear and high-dimensional contexts, deep learning-based methodologies have been proposed. The derivation of Granger causality within neural networks is based on the following definitions [180].

Definition 3.2 (Nonlinear GC between time-series). Given d time-series $\mathbf{X} = (\mathbf{x}_1^{1:T}, \ldots, \mathbf{x}_d^{1:T})$ across T time points and a non-linear function g_j, $\mathbf{x}_j^{t+1} = g_j(\mathbf{x}_1^{1:t}, \ldots, \mathbf{x}_d^{1:t}) + \mathbf{u}_j^{t+1}$, where $\mathbf{u}_j^{t+1}$ denotes independent noise. Time-series $\mathbf{x}_i$ is Granger non-causal for time-series $\mathbf{x}_j$ if for all $(\mathbf{x}_1^{1:t}, \ldots, \mathbf{x}_d^{1:t})$ and all $\mathbf{x}_i^{1:t} \neq \mathbf{x'}_i^{1:t}$,

$$g_j\left(\mathbf{x}_1^{1:t}, \ldots, \mathbf{x}_i^{1:t}, \ldots, \mathbf{x}_d^{1:t}\right) = g_j\left(\mathbf{x}_1^{1:t}, \ldots, \mathbf{x'}_i^{1:t}, \ldots, \mathbf{x}_d^{1:t}\right). \tag{11}$$

In recent years, emerging works have employed neural networks in Granger causal discovery.

(1) ***NGC and its Variants.*** Tank et al. [180] propose ***Neural Granger Causality (NGC)***, which leverages **component-wise MLP** (**cMLP**) or **component-wise LSTM** (**cLSTM**) with sparse input layer weights to infer nonlinear Granger causality. To be specific, in cMLP, each nonlinear output g_j is modeled with a separate MLP to distinguish the effects from inputs to outputs. With penalization, the input matrix of the first layer provides information for the selection of Granger causality. In the first layer of $g_j(\cdot)$, we have $h_1^t = \sigma(\sum_{k=1}^{\tau} W_1^k \mathbf{x}^{t-k} + b_1)$. If the ith column of weight matrix W_1^k contains zeros for all time lag k, then time-series $\mathbf{x}_i$ does not Granger-cause time-series $\mathbf{x}_j$. The objective function consists of a data-fitting term and a sparse-inducing term:

$$\text{min}_{\mathbf{W}} \sum_{t=\tau}^{T} \left(x_j^t - g_j(x_{(t-1):(t-\tau)})\right) + \lambda \sum_{i=1}^{d} R((W_1)_{:i}), \tag{12}$$

where sparse inducing penalty $R(\cdot)$ is implemented through group lasso penalty, which extracts causal relations without requiring precise lag specification. As for cLSTM, it sidesteps the lag selection problem and the Granger causal information is inferred from the input weight of LSTM. In ***economy Statistical Recurrent Units (eSRU)***, Khanna and Tan [99] extend cLTSM's ability in data scarcity scenarios based on statistical recurrent units.

(2) ***NN-GC and its variants.*** Montalto et al. [133] propose ***neural networks Granger causality (NN-GC)***, which is a feature selection procedure to identify the significant Granger causes in the MLP model. By greedily adding lagged components of predictor variable as input, an MLP is updated iteratively. A predictor variable is claimed to be a significant Granger cause of the target variable if at least one of its lagged components is considered as the input of neural networks when the procedure is terminated. In ***RNN-GC*** [190], the feature selection procedure is extended by replacing MLPs in *NN-GC* with gated RNN models, However, as this technique requires training and comparing many candidate models, it's costly in high-dimensional settings.

(3) ***GVAR***. Marcinkevics and Vogt [126] propose the ***generalized vector autoregression (GVAR)*** model for better interpretability in Granger causal discovery. It's based on a variant of self-explaining neural networks [3]. The self-explaining neural networks are inherently interpretable models motivated by restricted properties, and follow the form: $f(\mathbf{x}) = g(\theta(\mathbf{x})_1 h(\mathbf{x})_1, \ldots, \theta(\mathbf{x})_k h(\mathbf{x})_k)$, where $g(\cdot)$ and $\mathbf{h}(\mathbf{x})$ denote a link function and the interpretable basis concepts, respectively. Combined with the VAR model for MTS, the *GVAR* model is given by $\mathbf{x}^t = \sum_{l=1}^{\tau} \Psi_{\theta_l}(\mathbf{x}^{t-l})\mathbf{x}^{t-l} + \mathbf{u}^t$, where $\mathbf{u}^t$ denotes the additive noise term and $\Psi_{\theta_l} : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ is a neural network parameterized by θ_l, of which the output is the matrix corresponding to the strength of influence. In detail, the strength of influence $x_i^{t-l} \longrightarrow x_j^t$ is measured by the (i, j)-component of $\Psi_{\theta_l}(\mathbf{x}^{t-l})$. The loss function consists of three terms: the MSE loss, a sparsity-inducing regularization, and the smooth penalty. And variability of signs $(+/-)$ over time can also be captured.

(4) ***SCGL***. To infer Granger causality in high-dimensional settings, Xu et al. [193] propose the ***scalable causal graph learning (SCGL)*** framework based on matrix factorization and low-rank approximation. *SCGL* deconstructs data nonlinearity into two types (i.e., univariate-level and multivariate-level nonlinearity), which are modeled separately. The key idea of *SCGL* is that learning the full size of the adjacency matrix $A \in \mathbb{R}^{d \times d}$ would be unscalable when the size of variables d is quite large. In practice, the relationship of variables is low-rank in hidden space as in [32]. Therefore, *SCGL* approximates A via a k-rank decomposition, where $k < d$. The low-rank approximation reduces the noise influence in causal discovery and provides interpretability in downstream time-series analysis [85].

(5) ***TCDF***. Nauta et al. [135] propose the ***temporal causal discovery framework (TCDF)***, which utilizes attention-based dilated CNN to infer Granger causality. *TCDF* contains d independent attention-based CNNs for different target variables $(X_1, \ldots, X_d)$. For each target variable, a neural network is proposed to derive prediction, attention scores and kernel weights. Intuitively, a high attention score on X_i while forecasting X_j indicates the former contains prediction information toward the latter. *TCDF* evaluates variable importance and identifies significant causal links based on permutation. It can discover self-causation and time delays between cause and effect. Besides, by assuming that the bidirectional causal relations can not be instantaneous, it can also detect the presence of hidden confounders with equal delays.

(6) ***InGRA***. In ***inductive Granger causal modeling (InGRA)***, Chu et al. [34] combine the Granger causal attention in [161] with prototype learning to infer Granger causality in heterogeneous MTS data. To be specific, the Granger causal attention mechanism is first leveraged to compute contributions of each variable toward prediction. Due to data limitation, learned Granger causal attention is not robust enough. *InGRA* then leverages prototype learning, of which the key idea is to compute the similarities of new samples to prototypical cases, to detect common causal structures.

(7) ***ACD***. Löwe et al. [122] propose ***amortized causal discovery (ACD)***, which aims at training a model to infer Granger causal structures across different samples with various causal structures. Its implicit assumption is that different samples share similar dynamics. *ACD* is an encoder-decoder framework. The encoder function is defined to infer Granger causal relations of the input sample, and the decoder function forecasts the next time-step given the learned causal relations. As a result, the causal relations of previous unseen samples can be inferred without refitting the model.

(8) ***CR-VAE***. Li et al. [109] propose ***causal recurrent variational autoencoder (CR-VAE)*** where a generative model incorporates Granger causal learning into the data generation process. By preventing encoding future information before decoding, the encoder of *CR-VAE* obeys the principle of Granger causality An error-compensation module is leveraged to capture instantaneous effects. *CR-VAE* is not only able to infer Granger causality but also conduct the data-generating process in a transparent manner benefiting from the learned causal matrix.

3.5 Other Methods

The aforementioned four categories have been the primary focus of extensive research efforts in the field of MTS-CD. To provide a comprehensive overview, this section introduces four additional types of methods that are distinct from the aforementioned approaches. These include (i) *methods based on transfer entropy*, (ii) *methods based on differential equations*, (iii) *methods based on nonlinear state-space models*, and (iv) *methods based on logic formulas*.

3.5.1 Methods based on Transfer Entropy. Causality from time-series can be quantified with transfer entropy [160], a metric that measures the flow of information between two dynamic processes. For example, the transfer entropy from i to j (with time lag) can be expressed as

$$\mathrm{TE}\big(X_i^t \rightarrow X_j^{t+1}\big) = h\big(X_j^{t+1}|X_j^t\big) - h\big(X_j^{t+1}|X_j^t, X_i^t\big), \tag{13}$$

where $h(.|.)$ is the conditional entropy. Transfer entropy is capable of capturing nonlinear relationships without the need for specific models. Although the original definition of transfer entropy pertains to bivariate settings, the extension of this concept to multivariate scenarios is an area of interest, such as [6, 156, 176].

For Gaussian variables, Granger causality and transfer entropy are demonstrated to be equivalent [11]. To a certain extent, this family of methods, characterized by its model-free nature, shares commonalities with some constraint-based [7, 177] and Granger causality-based [101] approaches.

3.5.2 Methods based on Differential Equations. This particular family of methods is tailored for MTS systems that can be effectively represented by differential equations. The connection between differential equations and SCMs has been extensively explored in previous literature [143, 154]. In the context of discrete time, Voortman et al. [186] propose ***difference-based causality learner (DBCL)***, which can be regarded as a constrained version of dynamic SCMs. In the case of continuous time, theoretical efforts have been made to establish a causal interpretation of dynamic systems using both **Ordinary Differential Equations** (**ODEs**) [154] and **Stochastic Differential Equations** (**SDEs**) [74, 132].

More recently, ***neural graphical model (NGM)*** [13] has been proposed to model dynamic causal systems where the MTS data is irregularly sampled based on **Neural Ordinary Differential Equations** (**neural-ODE**). To be specific, the underlying causal system of interest can be represented as a dynamic structural model as follows:

$$d\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t))dt + d\mathbf{w}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \quad t \in [0, T], \tag{14}$$

where $\mathbf{w}(t)$ is a d-dimensional standard Brownian motion, $\mathbf{x}_0$ is a Gaussian random variable independent of $\mathbf{w}(t)$, and the function $\mathbf{f}$ describes the causal graph $\mathcal{G}$. *NGM* is a learning algorithm based on penalized neural-ODE. It's suited to irregularly sampled MTS.

3.5.3 Methods based on Nonlinear State-space. This family of methods is well-suited for situations characterized by the dynamic coupling of nonlinear systems. Sugihara et al. [175] propose ***convergent cross mapping (CCM)***, which infers causality based on Takens' state-space theory [179]. In detail, given two time-series $\mathbf{x}_1^t$ and $\mathbf{x}_2^t$, the attractor manifolds $\mathcal{M}_{x_1}, \mathcal{M}_{x_2}$ are first reconstructed. Secondly, causality can be detected by measuring the correspondence between $\mathcal{M}_{x_1}$ and $\mathcal{M}_{x_2}$, by testing whether one manifold preserves every local neighborhood defined in the other. Recent advancements [21] have been made to enhance *CCM* in challenging scenarios.

3.6 Discussions and Guidelines

From a practitioner's perspective, while it may initially appear challenging to achieve optimal model selection given the wide array of MTS-CD algorithms discussed above, it is possible to incrementally identify suitable solutions. This can be accomplished by systematically considering the characteristics of the available data, the objectives of downstream analyses, and the data volume [159]. In our context, the frameworks and insights summarized in Table 2 offer guidance in navigating these complexities.

To commence, it is essential to first consider the configurations of temporal data. It is crucial to distinguish between the problem of causal discovery from multivariate time-series (MTS-CD), discussed in this Section, and that from event sequences (ES-CD), which will be addressed in the subsequent Section. Following this, attention should be directed toward the configuration of variable values (i.e., discrete or continuous in Table 2). Furthermore, practitioners should carefully evaluate whether the data is irregularly sampled with continuous time intervals or discretely sampled with uniform spacing. As this distinction can significantly impact the suitability of different methodological approaches The assumptions made and the form of causal conclusions desired are also crucial factors in model selection. Specifically, practitioners might select algorithms depending on whether they need to account for lagged dependencies and contemporaneous relationships among variables.

Moreover, the intent and nature of downstream analysis will lead to preferred solutions [159]. If downstream tasks require a focus on the (conditional) independence relations between variables, constraint-based and score-based methods may be more suitable candidates. For instance, in root cause analysis within AIOps applications, which will be detailed in Section 6, the prevailing state-of-the-art approaches [89, 129] are often constraint-based or score-based methods. This preference arises from the necessity to uncover dependencies among IT components to enhance failure localization. In contrast, for prediction tasks or the analysis of temporal patterns such as trends and seasonality, methods based on Granger causality [114] are generally more appropriate.

From a technical perspective, the features and issues inherent in the data significantly influence the selection of appropriate algorithms for causal discovery. We have categorized these features and issues under the open issues section in Table 2, including nonlinearity, hidden confounders, non-stationarity, heterogeneity, and interventions, among others. Possessing prior knowledge of these issues allows practitioners to select algorithms with the corresponding capabilities that are better suited to handle these issues. Despite advancements, these issues remain challenging for causal discovery from temporal data, which will be explored in greater detail in the open issues section of Section 6.

Beyond the practitioner's perspective, future research in developing innovative algorithms should consider the significant potential of hybrid methods that integrate ideas from different paradigms. These methods can harness the strengths of diverse methodologies to create more robust solutions. For instance, SCM-based and constraint-based approaches have demonstrated enhanced robustness [8, 23], particularly under scenarios where traditional assumptions, such as the adjacency faithfulness assumption, may be violated.

4 Causal Discovery from Event Sequences

In real-world scenarios, the temporal events predominantly occur at irregular intervals. Consequently, it becomes imperative to model these asynchronous and irregular data in ES-CD. In line with the MTS-CD taxonomy, ES-CD methods can be classified into constraint-based, score-based, and Granger causality-based approaches. Among these categories, Granger causality-based methods have witnessed significant advancements owing to the inherent compatibility between Granger causality and several point processes. Therefore, in this chapter, we commence by introducing the application of multivariate point processes as a model for event sequence data. Subsequently, we delve into Granger causality methods based on Hawkes processes, Wold processes, and neural point processes. Lastly, we present an overview of other methods, including constraint-based and score-based approaches for ES-CD.

4.1 Multivariate Point Process

To capture high-dimensional event sequences, this section presents **Multivariate Point Processes (MPPs)** as a modeling framework. A high number of ES-CD methods revolves around aligning ES-CD with the task of MPP estimation. We initially introduce the *intensity function* and elucidate the rationale behind aligning these two tasks. Subsequently, we delve into the *log-likelihood function* associated with MPPs, which serves as a basis for estimation techniques.

A temporal point process refers to a stochastic or random process consisting of a series of binary events that transpire in continuous time [39]. MPPs are characterized by their high dimensionality, allowing for the inclusion of multiple types of events within the process. To be specific, the MPP with E types of events can be represented by E counting processes $\{N_e\}_{e=1}^{E}$, where $N_e = \{N_e(t)|t \in [0, T]\}$ and the occurring time of these events $\{t_1, t_2, \ldots, t_n | t_i \in [0, T]\}$ can be unevenly-distributed. The fundamental characterization of a point process is its conditional intensity function, which encapsulates the underlying pattern of the process. Specifically, the intensity function for a specific type u can be defined as the expected instantaneous rate of occurrence for events of type e, given the historical information:

$$\lambda_e(t) = \frac{\mathbb{E}[dN_e(t)|\mathcal{H}_t]}{dt}. \tag{15}$$

Here $\mathcal{E} = \{1, \ldots, E\}$ is the set of event types and $\mathcal{H}_t = \{(t_i, e_i)|t_i < t, e_i \in \mathcal{E}\}$ represents all types of events happened before time t.

In other words, $\lambda_e(t)\Delta t$ represents the instantaneous probability of type-e event's occurrence in the time window $[t, t + \Delta t]$, which aligns to causal strength when condition on other variables. Hence, the ES-CD problem can be formulated as the task of taking a collection of point processes, where each individual process represents a sequence of events, and generating a causal graph $\mathcal{G}$ constructed from these processes. Within the causal graph $\mathcal{G}$, each node corresponds to a point process, while each directed edge represents a causal interaction from one point process to another. By applying the chain rule, the likelihood function [88] for the joint distribution can be established:

$$L_0 \triangleq \sum_{j=1}^{n} lnf(t_j|e_j, \mathcal{H}_{t_j}) + \sum_{j=1}^{n} lnf(e_j|\mathcal{H}_{t_j}). \tag{16}$$

To infer the causal relationship between different events, the focus lies on the first term [88]. In the preceding discussion, we provided a brief introduction to MPPs and established likelihood functions to describe MPPs. Moving forward, our focus will shift toward a comprehensive exploration of detailed ES-CD methods.

4.2 Granger Causality-based Methods

In a manner akin to that observed in MTS, the causality of e_j-type events upon e_i is established through Granger causality, wherein the predictive utility of $e_j(t)|t < t_0$ in forecasting $e_i(t)$ is examined [51, 102]. When considering the specifications of the model, the methodologies can be classified into three categories: Hawkes process-based, Wold process-based, and neural point process-based approaches. Our review will encompass an analysis of their distinctive characteristics and delineate the scenarios in which they are best suited.

4.2.1 Hawkes Process-based Methods. As a distinct type of point process, we commence by presenting the fundamental aspects of the Hawkes process, encompassing the formulation of the intensity function and the inferential procedures employed to ascertain Granger causality based on the intensity function. Subsequently, we delve into a comprehensive exploration of the intricate methodologies associated with this domain.

The Hawkes process [76], classified as a point process, exhibits temporal dynamics characterized by either intensities that increase abruptly or decay gradually. Notably, this process adheres to a predetermined structure of the intensity function, denoted as

$$\lambda_{e_i}(t) = \mu_{e_i} + \sum_{e_j=1}^{E} \int_0^t \phi_{e_i e_j}(s) dN_{e_j}(t-s). \tag{17}$$

In this context, the term μ_{e_i} refers to the baseline intensity, which remains unaffected by endogenous events and therefore remains constant throughout the temporal domain. Conversely, the function $\phi_{e_i e_j}(s)$ serves as the impact function, quantifying the attenuation of the influence imparted by preceding type-e_j events on subsequent occurrences of type-e_i events. Essentially, the impact function encapsulates the internal intensity transfer from e_j to e_i. Given the resemblance observed in the definitions of ϕ and Granger causality, it is plausible to deduce Granger causality directly through the examination of ϕ. To be specific, the following proposition [51] holds:

$$e_j \text{ does not Granger-cause } e_i \Leftrightarrow \phi_{e_i e_j}(t) = 0, \forall t \in \mathbb{R},$$

where $\mathbb{R}$ indicates the set of all real numbers. Consequently, it becomes possible to formulate $\phi_{e_i e_j}(t)$ or its modified versions as a means to deduce Granger causality from event sequence data modeled by the Multivariate Hawkes process. Notably, numerous studies [2, 25, 27, 81, 88, 91, 93, 111, 194, 206, 214] have emerged within this domain in recent years.

(1) ***MLE-SGLP***. Xu et al. [194] propose ***MLE-based algorithm with Sparse-Group-Lasso and Pairwise (MLE-SGLP)*** *similarity constraints*, which learns Granger causality from Multivariate Hawkes processes with a maximum likelihood estimator and a sparse-group-lasso regularizer. Nonetheless, the *MLE-SGLP* may encounter limitations due to its elevated computational complexity and limited modeling capability. Parameterization strategies on intensity functions and regularization methods on likelihood functions can be utilized to augment the modeling capacity. While certain techniques such as Generalized Method of Moments and event sequence separation can be leveraged to reduce the computational complexity.

(2) ***NPHC***. Achab et al. [2] argue that causality in the Multivariate Hawkes process can be inferred without estimating the shape of the impact function and they propose ***Non Parametric Hawkes Cumulant (NPHC)***. Instead of the detailed form of the intensity function, *NPHC* infers Granger causality based on its integrals. *NPHC* has a competitive computational complexity compared with methods such as *MLE-SGLP*, This is attributed to *NPHC*'s innovative estimation technique, which focuses on the integrals of intensity functions rather than their explicit forms, thereby optimizing computational efficiency.

(3) L_0***Hawkes***. Idé et al. [88] claim that EM-based MLE algorithms with L_1-regularization cannot offer sparse solutions for Hawkes process mathematically. Hence, they propose L_0*Hawkes*, which is an L_0-regularized EM-based MLE algorithm to circumvent this problem. To be specific, the L_0-norm $||A||_0$ indicates counts of non-zero entries in matrix $\mathbf{A}$. This model exhibits a substantial enhancement in its modeling capacity. Detailed theoretical analysis can be found in [88].

(4) ***GC-nsHP***. Chen et al. [27] propose ***Granger Causality for non-stationary Hawkes Process (GC-nsHP)***. It first assumes that a non-stationary Hawkes process can be approximated by utilizing a mixture of non-overlapping, and stationary processes. Subsequently, a dynamic programming-based algorithm is employed to partition the non-stationary process into multiple stationary sub-processes. At last, *GC-nsHP* leverages an EM-based algorithm to learn Granger causality in each sub-process. It demonstrates a substantial improvement in its computational complexity

(5) ***THP***. Cai et al. [25] propose ***Topological Hawkes Process (THP)*** to address the Granger causal discovery problem by incorporating prior knowledge of the underlying topological

relationships. For example, in telecommunication networks, alarms as events are propagated across different devices with an underlying topological structure.[1] *THP* perceive the conventional Hawkes process as a temporal convolution and subsequently expand it into the time-graph domain by incorporating graph convolutional networks. Subsequently, an optimization method combining an EM-based algorithm with hill-climbing is employed. *THP* could significantly improve the model capacity when information about the underlying topological structure is available. However, it also entails a relatively high computational complexity.

(6) ***MDLH***. Jalaldoust et al. [91] introduce the method *MDLH*, which employs **Minimum Description Length** (**MDL**) as the principle to frame the Granger causal problem as a model selection task. MDL, as a practical tool based on information compression, facilitates model selection. In the context of Granger causal discovery, *MDLH* leverages a Monte-Carlo method for inference. It is especially suitable for estimating Granger causalities from short event sequences.

(7) ***MMLH***. Hlaváčková-Schindler et al. [81] propose the method *MMLH*, in which the Granger causality among dimensions of multivariate Hawkes processes are seen as a variable selection problem and is approached by minimum message length principle. Similarly as *MDLH*, it is suitable for estimating Granger causalities from short event sequences.

4.2.2 *Wold Process-based Methods.*

Wold process is a class of point process defined in terms of a Markovian joint distribution of inter-event times, and is used to model causality in event sequences with reduced complexity [42, 53]. In this part, we will begin by providing an overview of the Wold process, highlighting its differentiation from the Hawkes process. Subsequently, we will delve into a comprehensive examination of the methodologies employed for Granger causal discovery.

Contrasting with the Hawkes process, where the intensity function relies on the entire history of past events, the Wold process exhibits a Markov chain of finite memory for its inter-event time intervals. To be specific, we denote $\delta_i = t_i - t_{i-1}$ as the waiting time for the ith event from the occurrence of the $(i-1)$-th event. Wold Processes are constructed based on the fundamental assumption that the current waiting time δ_i is solely dependent on the closest preceding waiting time δ_{i-1}. A commonly employed representation (Busca model [38]) of the intensity function for the Multivariate Wold process can be expressed as

$$\lambda_{e_i}(t) = \mu_{e_i} + \sum_{e_j \in E} \frac{\alpha_{e_i e_j}}{\beta_{e_j} + \Delta_{e_i e_j}(t)}, \tag{18}$$

where μ_{e_i} denotes the Poisson rate that characterizes the exogenous events, and $\sum_{e_j \in E} \frac{\alpha_{e_i e_j}}{\beta_{e_j} + \Delta_{e_i e_j}(t)}$ represents the endogenous Wold rate, which is regulated by the time interval and the strength of interaction. These distinctive attributes contribute to a lower time complexity in estimating the Wold process. Moreover, it's more suitable in contexts such as communication dynamics [53].

In recent years, studies [42, 53] have focused on inferring Granger causality from the Multivariate Wold process.

(1) ***Granger-Busca***. Figueiredo et al. [42] introduce *Granger-Busca*, the pioneering approach to learning Granger causality from event sequences modeled by Multivariate Wold processes. *Granger-Busca* utilizes a **Markov Chain Monte Carlo** (**MCMC**) sampling algorithm for parameter optimization, and Granger causality between e_i and e_j is established when $\alpha_{e_i e_j} \neq 0$ in Equation (18). Leveraging the inherent characteristics of the Wold process, *Granger-Busca* demonstrates significantly improved computational efficiency compared to Hawkes process-based methods. However, *Granger-Busca* imposes several constraints on the structure. For instance, it necessitates that each node possesses at least one outgoing edge.

[1] https://neurips.cc/virtual/2023/competition/66582

Table 3. A Comparison between Granger Causality Discovery Methods on Event Sequences

Model	Idea	Model Capacity	Complexity	Data Volume	Suitable when
MLE-SGLP [194]	Hawkes+MLE	Low	High	Low	–
NPHC [2]	Hawkes+GMM	Low	Low	High	–
L_0Hawkes [88]	Hawkes+L_0 penalty	High	High	Low	–
GC-nsHP [27]	Hawkes+ES separation	Low	Low	High	Non-stationary HP
THP [25]	Hawkes+Topological graph	High	High	Low	Underlying topology
MDLH [91]	Hawkes+MDL principle	High	–	High	Short event sequences
MMLH [81]	Hawkes+MML principle	High	–	High	Short event sequences
Granger-Busca [42]	Wold+MCMC	Low	Low	High	–
Var-Wold [53]	Wold+Variational Inference	Low	Low	High	–
CAUSE [212]	Encoder-decoder+Attribution	Medium	High	High	Inhibitive causality
TNPAR [24]	Neural Poisson+Amortized infer.	High	High	High	Underlying topology

(2) ***Var-Wold***. In order to alleviate the assumptions made by *Granger-Busca*, Etesami et al. [53] introduce *Var-Wold*, a novel framework that utilizes a Bayesian approach (variational inference) to optimize parameters in multivariate Wold processes. Furthermore, owing to the Markovian nature of the intensity function, the Bayesian approach overcomes the issue of long memory present in the Hawkes process.

4.2.3 Neural Point Process-based Methods. With the rapid advancement of neural networks, **Neural Point Processes (NPPs)** have gained traction in modeling event sequences and inferring causal relationships. In contrast to statistical point processes such as Hawkes processes and Wold processes, NPPs harness the learning capabilities of neural networks to effectively model sequences, often surpassing statistical point processes in terms of predictive performance [24]. At the core of these NPP algorithms lies the concept of employing neural networks to estimate the intensity function $\lambda_e(t)$. Specifically, they encode the event sequence into a hidden state, capturing its underlying features, and subsequently employ decoders to infer the future intensity function.

Several studies [24, 212] have emerged within this domain in recent years.

(1) ***CAUSE***. Zhang et al. [212] propose *CAUSE*, which infers Granger causality from NPPs. As NPPs are less interpretable and unable to directly reveal Granger causality, attribution methods are leveraged. This method can also measure the inhibitive causality, in which an event has the ability to inhibit or suppress the occurrence of subsequent events, and the magnitude of the causality.

(2) ***TNPAR***. Cai et al. [24] introduce ***Topological Neural Poisson Autoregressive (TNPAR)*** model, which integrates prior knowledge of the underlying topological relationships to learn Granger causality, resembling the approach employed by *THP*. *TNPAR* comprises two distinct stages: (i) the generation stage, and (ii) the inference stage. During the generation stage, to model dynamics entailed in event sequences, *TNPAR* leverages a modified version of the neural Poisson process. Then, an amortized inference procedure is employed to estimate causality.

Table 3 is a comparison between all Granger causality discovery methods on event sequences. Different approaches' ideas, model capacities, complexities, data volumes, and the scenarios where they can be most effectively utilized are being compared.

Furthermore, there have been studies [167, 168] focusing on the estimation of influences between events within NPPs, akin to the notion of Granger causality within the context of NPPs.

4.3 Other Methods

In addition to the predominant utilization of Granger causality-based methods in ES-CD, there exist other endeavors [1, 18, 105]. Specifically, constraint-based and score-based methods are also employed, which align with those employed in MTS-CD. Both the constraint-based methods and the score-based methods are based on **Graph Event Models** (**GEM**), which are graphical representations for Multivariate Point Processes. In Specific, A Graphical Event Model is a pair $< \mathcal{G}, \Theta >$. $\mathcal{G}$ is a graph in which each vertex represents a point process, or in other words, a flow of a particular type of event. The directed edges represent the potential effects from a point process to another point process. Θ is the parameterization set in which each element parameterizes a particular intensity function of a point process. This section will provide a comprehensive review of these two categories of methods.

4.3.1 Constraint-based Methods. Constraint-based methods rely on the assessment of independence between processes, akin to the examination of conditional independence and d-separation in MTS-CD. Corresponding notions [45] and procedures can be identified in constraint-based methods [1, 17] applied to event sequences.

(1) ***μ-PC***. Absar and Zhang [1] propose a constraint-based approach called *μ-PC*, which leverages μ-separation [131] and extends *PC*. μ-separation offers a formal graphical depiction of statistical relationships within temporal data. To handle event sequence data, a novel technique for testing conditional independence is introduced, leveraging the **Recurrent Marked Temporal Point Process** (**RMTPP**) model [48].

(2) ***CB-PGEM***. Bhattacharjya et al. [17] propose ***Constraint-Based Proximal Graphical Event Model (CB-PGEM)***, which introduces process independence tests for causal proximal graphical event model [18] and uses constraint-based structure discovery algorithms (e.g., max-min parents algorithm) for Bayesian networks to deploy these tests.

4.3.2 Score-based Methods. Score-based ES-CD methods revolve around the inference of causality through the optimization of score functions, a domain that has witnessed several studies [15, 16, 18].

(1) ***SB-PGEM** and its Variants*. Bhattacharjya et al. [18] propose ***Score-Based Proximal Graphical Event Model (SB-PGEM)*** which employs a score-based optimization approach to estimate the model parameters. The authors utilize BIC as the score to search for optimal time windows and parent sets for each event type. In a more recent work, Bhattacharjya et al. [16] extend *SB-PGEM* while introducing the Bayesian Gamma score to be the score function to address the challenges of limited data by incorporating background knowledge into the modeling process.

(2) ***CIR***. Disregarding the conditional causal relationship, Bhattacharjya et al. [15] introduce a collection of scores, namely conditional intensity-based scores, along with associated algorithms to estimate the cause-effect associations between event pairs derived from extensive event datasets.

5 Resources

In this section, we summarize available resources of causal discovery for temporal data, including evaluation metrics and public datasets. Due to the space limit, we just mention the outlines here. More details about these resources can be found in the supplementary materials.

Evaluation Metrics. Commonly employed evaluation metrics include **Accuracy** (**ACC**), **Area Under the Receiver Operator Curve** (**AUROC**), and **Structural Hamming Distance** (**SHD**).

Datasets. We provide an overview of publicly available datasets suitable for evaluating both MTS-CD and ES-CD methodologies. Specifically, the datasets employed for MTS-CD encompass (i) *Lorenz-96* [121], (ii) *CMU-MoCap* [180], (iii) *DREAM-3* [148], (iv) *BOLD* [170], (v) *Financial* [104], (vi) *MIMIC-III* [95], (vii) *CIPCaD-Bench* [128], (viii) *TECDI* [110], and (ix) *CausalTime* [29]. While the datasets utilized for ES-CD comprise (i) *MemeTracker* [2], (ii) *IPTV* [123], (iii) *Grid* [88], and (iv) *G-7 bonds* [44].

Table 4. The Summary of Representative Applications, where SOTA Indicates the State-of-the-Art Temporal Causal Discovery Methods

Application	Data Type	Issue	SOTA	Example
AIOps	MTS: metrics, KPIs, and so on	Temporal Pattern	GC(NGC), CB(PCMCI)	[114], [129]
		Latent Interventions	CB(ψ-PC), SB	[89], [188]
	ES: alarms, logs, and so on	Topological Structure	SB, GC	[189], [215], [24]
Climatology	MTS: climate indices, and so on	Heterogeneity	GC	[80], [183]
		Non-stationarity	CB(Regime-PCMCI)	[97]
		Knowledge Incorporation	SB	[127]
		Scalability	CB(girded)	[77], [182], [19]
Healthcare	MTS: vital signs, ECGs, fMRIs, and so on	Irregularly Sampling	SB(CTBN), GC(NGC)	[117], [30], [28]
		Non-stationarity	GC(NGC), CB	[92], [54]
		Temporal Pattern	GC(NGC)	[109]
	ES: EHRs, and so on	Scalability	GC	[191]

6 Applications, Open Issues, and New Perspectives

6.1 Applications

Causal discovery from temporal data finds utility in diverse real-world applications, including scientific research and industrial applications. Applications in scientific research encompass a wide range of domains, including climatology, healthcare. The causal findings obtained through these analyses often serve as preliminary hypotheses and pivotal starting points for subsequent investigations [195]. In the realm of industrial applications, temporal causal discovery finds extensive usage in tasks such as anomaly detection [112], root cause analysis [62], spatial-temporal analysis [73, 198], and more. As an enabling tool, causal discovery assumes a supportive role within a multi-stage approach in an industrial context [62]. In Table 4, we delineate three key application domains: **Artificial Intelligence for IT Operations** (**AIOps**), climatology, and healthcare. For each application, the table provides a detailed overview of the relevant data types, the corresponding challenges, and the state-of-the-art methodologies employed. Further insights into applications of causal discovery from temporal data can be found in the supplementary material.

6.2 Open Issues

Causal discovery from temporal data is a prominent research area in the fields of data mining and machine learning. Numerous publications about this topic have emerged in top conferences and journals. Despite the substantial progress made, several challenges persist and warrant further investigation, including non-stationarity, data heterogeneity, data subsampling, hidden confounder, and collaborating with expert knowledge.

Non-stationarity. Non-stationarity refers to the phenomenon where the underlying causal mechanisms within a system change over time. This can manifest as variations in the causal structure (i.e., structural non-stationarity), changes in the strength of causal effects, or shifts in the characteristics of noise terms (i.e., parameter non-stationarity) [58, 151, 184, 211]. For example, in the financial sector, market conditions and policies are continually evolving, resulting in non-stationary financial time-series data [84, 184]. Similarly, in gene regulatory networks, the expression of genes and their interactions can change in response to environmental factors, leading to non-stationarity in biological data [46]. According to whether the changepoints and number of regimes are given, there exist diverse non-stationary settings [151, 184] in reality which vary in complexity from easy to difficult, including **Known Number and Known Times of Transitions** (**KNKT**), **Known Number But Unknown Times** (**KNUT**) of Transitions, and **Unknown Number and Unknown Times** (**UNUT**) of Transitions. Some existing works have addressed non-stationarity from various angles. For example, non-stationary DBNs [151] and non-stationary CTBNs [184] are defined and inferred under diverse settings mentioned above. And some approaches utilize surrogate variables [84, 211] to account for changing causal relations, while others employ evolutionary spectral analysis [47] for slowly varying processes. Despite these efforts, most methods rely on strict model assumptions, which can limit their applicability and effectiveness in practice. This reliance on assumptions often leaves non-stationarity as a persistently unresolved issue within the field, necessitating further research to develop more flexible and robust methods.

Data Heterogeneity. Another tangible challenge encountered across various real-world applications is data heterogeneity, which indicates divergence in the data-generating processes or causal mechanisms among different samples within a dataset [12, 34, 61]. For instance, in healthcare, patient data may exhibit heterogeneity due to differences in demographics, or treatment responses, leading to distinct causal pathways in disease progression or therapeutic effectiveness. Although several remedies are proposed such as [12, 34, 64, 82, 122], this disparity still significantly complicates the task of identifying consistent and generalizable causal relationships. This issue extends beyond the temporal dimension, and it's a fundamental challenge in the field of causal discovery.

Sampling and Missing Data. Temporal data, especially time-series, may be collected at a lower rate than that of the underlying causal process due to the difficulties in data sampling. Without considering this property, causal discovery algorithms may generate spurious or missing causal relations. Although several remarks and approaches [40, 65, 181, 203] are proposed for this issue, the subsampling problem in temporal data has not been well addressed. Another fundamental issue in temporal data is irregular sampling. In [117], CTBN is studied in the clinical time-series which is irregularly sampled. And this study provides insight that CTBN outperforms its discrete time interval counterpart (i.e., DBN) in more realistic settings where data is not missing completely at random. To be specific, observations in healthcare may be affected by many factors in reality, e.g., whether the patient feels ill [117]. Recently, several initiatives [13, 28, 30] have also emerged within this pioneering field.

Hidden Confounders. In practice, we are often met with cases where causal sufficiency is violated, i.e., there exist hidden confounders. This challenging setting may lead to incorrect causal relations [59]. As summarized in Table 2, most temporal causal discovery approaches cannot handle hidden confounders in a straightforward way. In some recent works, hidden confounders are either modeled by applying a structural bias [122] or termed as causal representation learning techniques [116]. However, all these methods have strong assumptions about the influence mechanism of confounders. How to discover the causal relationships with hidden confounders remains an open issue. The issue of hidden confounders is a fundamental challenge in causal discovery, extending beyond the confines of temporal data, and represents a particularly complex problem.

Collaborating with Expert Knowledge. Expert knowledge can boost causal discovery in practice. The approaches of fusing expert knowledge can be grouped into three types [103]: (i) *Soft constraints*: the learning process can be influenced by the knowledge [139]. (ii) *Hard constraints*: the learned structure must conform to the enforced requirements. In [9], hard constraints are leveraged in structure learning with a time-dependant exposure. Studies in [178] add prior knowledge forbidding the existence of intra-slice dependencies. (iii) *Interactive learning*: the human input is leveraged in the learning process [127, 130]. Nevertheless, how to collaborate with expert knowledge needs to be further explored.

6.3 New Perspectives

6.3.1 A Unified Framework for Temporal Causal Discovery. Given realistic demands and the reviewed methods, we assert the necessity and feasibility of a unified causal discovery framework for both MTS and ES data. In real-world scenarios, heterogeneous temporal data from diverse sources are encountered, such as MTS data from sensors and ES data extracted from text posts. Relying solely on either MTS-CD or ES-CD may introduce confounding factors. Existing MTS-CD and ES-CD methods discussed in Section 3 and Section 4 share a similar taxonomy and exhibit fundamental consistency, paving the way for a unified framework in future temporal causal discovery research.

6.3.2 New Problems of Temporal Causal Discovery. The concept of temporal causal discovery, in its various forms, can empower numerous facets of machine learning and data mining.

We emphasize three new problems of temporal causal discovery that have not been well studied. (i) *Online temporal causal discovery.* Different from previous methods which train distinct models for individual samples, a comprehensive framework for global causal discovery is trained to accommodate individuals exhibiting diverse causal structures [34, 122]. They effectively utilize information from extensive datasets, enabling the inference of causal relationships for newly encountered individuals. (ii) *Supervised temporal causal discovery.* Distinct from the conventional frameworks, the inference process is treated as a black box to learn the mapping from sample data to causal graph structures [14, 146], which aligns with the paradigm with most artificial intelligence tasks. (iii) *Causal representation learning for temporal data.* The aforementioned causal discovery methods focus on inferring relations between observed variables, or start from the premise that the causal variables are given beforehand. As a generalization of temporal causal discovery from observed variables, there has recently been a growing interest in temporal causal representation learning [116, 202], which aims at the learning representation of causal factors in an underlying dynamic system. It estimates latent causal variable graphs from observations.

6.3.3 Large Language Models for Causal Discovery. Recently, we have witnessed great breakthroughs of **Large Language Models (LLMs)** in many domains[113, 118]. It reveals the synergistic potentials between causal discovery and LLMs in two aspects: (i) How do LLMs affect causal discovery research? (ii) How can causal methods benefit the advances of LLMs?

Owing to LLMs, extant causal discovery methods can be augmented by prior knowledge extracted from a large-scale human corpus dataset. In [100], Kıcıman et al. find that GPT-3.5/4 can get competitive results in the causal discovery task by merely leveraging label information. However, researchers [208] argue that current LLMs relying on human knowledge may not be enough to offer new insights that are beyond limitations of human intuition. Therefore, recent researches [120] reveal that LLMs open a door for the collaborative improvement between data-driven insights from causal discovery algorithms and causal knowledge from human corpus dataset.

Owing to causal research, opportunities are presented to enhance current LLMs in two key aspects. Firstly, it fosters the trustworthiness of LLMs [119, 208] by eliminating bias and generating interpretable results. Secondly, by integrating causal-inspired approaches, LLMs' capacity for causal reasoning [94, 208] can be further advanced.

7 Conclusion

In this survey, we systematically review causal discovery methods for temporal data and provide some new perspectives. According to the nature and structure of temporal data, we divide recent research works into two groups, i.e., causal discovery from multivariate time-series (MTS-CD) and causal discovery from event sequence (ES-CD). Moreover, a method taxonomy according to the different definitions of causalities is proposed to reorganize the techniques in both MTS-CD and ES-CD. Compared with other surveys, this is the first work that summarizes the causal discovery methods for MTS and ES systematically. For broader applications, we identified several open issues that require urgent attention from researchers. Moreover, new perspectives on temporal causal discovery are presented. We hope this survey can empower machine learning researchers to solve causal problems and shed light on the usage of causal discovery techniques.

Acknowledgments

We thank Lun Du, Wei Chen, Jin Wang, Xinxin Fan, Fei Wang, Zezhi Shao, Hao Sun, and Haolan Wang for valuable advice. We are grateful to editors and reviewers for constructive feedback.

References

[1] 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*. 2807–2811.

[2] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. 2017. Uncovering causality from multivariate hawkes integrated cumulants. In *Proceedings of the ICML*. 1–10.

[3] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*. 7786–7795.

[4] Andrew Arnold, Yan Liu, and Naoki Abe. 2007. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 66–75.

[5] Antonin Arsac, Aurore Lomet, and Jean-Philippe Poli. 2023. Causal discovery for time series with constraint-based model and PMIME measure. In *When Causal Inference meets Statistical Analysis*. Paris, France.

[6] Charles K. Assaad, Emilie Devijver, and Éric Gaussier. 2022. Discovery of extended summary graphs in time series. In *Proceedings of the Uncertainty in Artificial Intelligence*. 96–106.

[7] Charles K. Assaad, Emilie Devijver, and Éric Gaussier. 2022. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research* 73 (2022), 767–819.

[8] Charles K. Assaad, Emilie Devijver, Éric Gaussier, and Ali Aït-Bachir. 2021. A mixed noise and constraint-based approach to causal inference in time series. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 1317, 2021, Proceedings, Part I 21*. 453–468.

[9] Vahé Asvatourian, Philippe Leray, Stefan Michiels, and Emilie Lanoy. 2020. Integrating expert's knowledge constraint of time dependent exposures in structure learning for bayesian networks. *Artificial Intelligence in Medicine* 107 (2020), 101874.

[10] Zhidong Bai, Wing-Keung Wong, and Bingzhi Zhang. 2010. Multivariate linear and nonlinear causality tests. *Mathematics and Computers in Simulation* 81, 1 (2010), 5–17.

[11] Lionel Barnett, Adam B. Barrett, and Anil K. Seth. 2009. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters* 103, 23 (2009), 238701.

[12] Sahar Behzadi, Katerina Hlavácková-Schindler, and Claudia Plant. 2019. Granger causality for heterogeneous processes. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part III 23*. 463–475.

[13] Alexis Bellot, Kim Branson, and Mihaela van der Schaar. 2022. Neural graphical modelling in continuous-time: Consistency guarantees and algorithms. In *Proceedings of the International Conference on Learning Representations*.

[14] Danilo Benozzo, Emanuele Olivetti, and Paolo Avesani. 2017. Supervised estimation of granger-based causality between time series. *Frontiers in Neuroinformatics* 11 (2017). https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2017.00068/full

[15] Debarun Bhattacharjya, Tian Gao, Nicholas Mattei, and Dharmashankar Subramanian. 2020. Cause-effect association between event pairs in event datasets. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*. ijcai.org, 1202–1208.

[16] Debarun Bhattacharjya, Tian Gao, Dharmashankar Subramanian, and Xiao Shou. 2023. Score-based learning of graphical event models with background knowledge augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 12189–12197.

[17] Debarun Bhattacharjya, Karthikeyan Shanmugam, Tian Gao, and D. Subramanian. 2022. Process independence testing in proximal graphical event models. In *Proceedings of the Conference on Causal Learning and Reasoning*. 144–161.

[18] Debarun Bhattacharjya, Dharmashankar Subramanian, and Tian Gao. 2018. Proximal graphical event models. In *Proceedings of the Advances in Neural Information Processing Systems*. 8147–8156.

[19] A. Böhnisch, E. Felsche, and R. Ludwig. 2023. European heatwave tracks: using causal discovery to detect recurring pathways in a single-regional climate model large ensemble. *Environmental Research Letters* 18, 1 (2023), 014038.

[20] Alessandro Bregoli, Marco Scutari, and Fabio Stella. 2021. A constraint-based algorithm for the structural learning of continuous-time bayesian networks. *International Journal of Approximate Reasoning* 138 (2021), 105–122.

[21] Edward De Brouwer, Adam Arany, Jaak Simm, and Yves Moreau. 2021. Latent convergent cross mapping. In *Proceedings of the International Conference on Learning Representations*.

[22] Kenneth P. Burnham and David R. Anderson. 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research* 33, 2 (2004), 261–304.

[23] Daria Bystrova, Charles K. Assaad, Julyan Arbel, Emilie Devijver, Éric Gaussier, and Wilfried Thuiller. 2024. Causal discovery from time series with hybrids of constraint-based and noise-based algorithms. *Transactions on Machine Learning Research Journal* 2024 (2024).

[24] Ruichu Cai, Yuequn Liu, Wei Chen, Jie Qiao, Yuguang Yan, Zijian Li, Keli Zhang, and Zhifeng Hao. 2024. TNPAR: Topological neural poisson auto-regressive model for learning granger causal structure from event sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 20491–20499.

[25] Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. 2024. THPs: Topological hawkes processes for learning causal structure on event sequences. *IEEE Transactions on Neural Networks and Learning Systems* 35, 1 (2024), 479–493.

[26] Luca Castri, Sariah Mghames, Marc Hanheide, and Nicola Bellotto. 2023. Enhancing causal discovery from robot sensor data in dynamic scenarios. In *CLeaR (Proceedings of Machine Learning Research, Vol. 213)*. PMLR, 243–258.

[27] Wei Chen, Jibin Chen, Ruichu Cai, Yuequn Liu, and Zhifeng Hao. 2022. Learning granger causality for non-stationary hawkes processes. *Neurocomputing* 468 (2022), 22–32.

[28] Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. 2024. CUTS+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 11525–11533.

[29] Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. 2024. CausalTime: Realistically generated time-series for benchmarking of causal discovery. In *Proceedings of the International Conference on Learning Representations*.

[30] Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. 2023. CUTS: Neural causal discovery from irregular time-series data. In *Proceedings of the International Conference on Learning Representations*.

[31] David Maxwell Chickering. 1995. Learning bayesian networks is NP-complete. In *Artificial intelligence and statistics V*. 121–130.

[32] Alessandro Chiuso and Gianluigi Pillonetto. 2012. A bayesian approach to sparse dynamic network identification. *Automatica* 48, 8 (2012), 1553–1565.

[33] Tianjiao Chu and Clark Glymour. 2008. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research* 9 (2008), 967–991.

[34] Yunfei Chu, Xiaowei Wang, Jianxin Ma, Kunyang Jia, Jingren Zhou, and Hongxia Yang. 2020. Inductive granger causal modeling for multivariate time series. In *Proceedings of the 2020 IEEE International Conference on Data Mining*. 972–977.

[35] Nevin Climenhaga, Lane DesAutels, and Grant Ramsey. 2021. Causal inference from noise. *Noûs* 55, 1 (2021), 152–170.

[36] Daniele Codecasa and Fabio Stella. 2014. Learning continuous time bayesian network classifiers. *International Journal of Approximate Reasoning* 55, 8 (2014), 1728–1746.

[37] Anthony C. Constantinou, Yang Liu, Kiattikun Chobtham, Zhigao Guo, and Neville Kenneth Kitson. 2021. Large-scale empirical validation of bayesian network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning* 131 (2021), 151–188.

[38] Rodrigo Augusto da Silva Alves, Renato Martins Assunção, and Pedro Olmo Stancioli Vaz de Melo. 2016. Burstiness scale: A parsimonious model for characterizing random series of events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1405–1414.

[39] Daryl J. Daley and David Vere-Jones. 2003. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer.

[40] David Danks and Sergey Plis. 2013. Learning causal structure from undersampled time series. In *JMLR: Workshop and Conference Proceedings (NIPS Workshop on Causality)*. 1–10.

[41] Cassio P. de Campos and Qiang Ji. 2011. Efficient structure learning of bayesian networks using constraints. *Journal of Machine Learning Research* 12 (2011), 663–689.

[42] Flavio V. D. de Figueiredo, Guilherme Resende Borges, Pedro O. S. Vaz de Melo, and Renato M. Assunção. 2018. Fast estimation of causal interactions using wold processes. In *Proceedings of the Advances in Neural Information Processing Systems*. 2975–2986.

[43] Kevin Debeire, Jakob Runge, Andreas Gerhardus, and Veronika Eyring. 2024. Bootstrap aggregation and confidence measures to improve time series causal discovery. In *CLeaR (Proceedings of Machine Learning Research, Vol. 236)*. PMLR, 979–1007.

[44] Mert Demirer, Francis X Diebold, Laura Liu, and Kamil Yilmaz. 2018. Estimating global bank network connectedness. *Journal of Applied Econometrics* 33, 1 (2018), 1–15.

[45] 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.

[46] Frank Dondelinger, Sophie Lèbre, and Dirk Husmeier. 2013. Non-homogeneous dynamic bayesian networks with bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Machine learning* 90, 2 (2013), 191–230.

[47] Kang Du and Yu Xiang. 2024. Causal inference from slowly varying nonstationary processes. *IEEE Trans. Signal Inf. Process. over Networks* 10 (2024), 403–416.

[48] 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.

[49] Tom Edinburgh, Stephen J. Eglen, and Ari Ercole. 2021. Causality indices for bivariate time series data: A comparative review of performance. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 31, 8 (2021), 083111.

[50] Michael Eichler. 2012. *Causal Inference in Time Series Analysis*. Wiley Online Library.

[51] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. 2017. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis* 38, 2 (2017), 225–242.

[52] Doris Entner and Patrik O. Hoyer. 2010. On causal discovery from time series data using FCI. *Probabilistic Graphical Models* (2010), 121–128.

[53] Jalal Etesami, William Trouleau, Negar Kiyavash, Matthias Grossglauser, and Patrick Thiran. 2021. A variational inference approach to learning multivariate wold processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 2044–2052.

[54] Muhammad Hasan Ferdous, Uzma Hasan, and Md. Osman Gani. 2023. CDANs: Temporal causal discovery from autocorrelated and non-stationary time series data. In *MLHC (Proceedings of Machine Learning Research, Vol. 219)*. Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary C. Lipton, Rajesh Ranganath, Iñigo Urteaga, and Serene Yeung (Eds.). PMLR, 186–207.

[55] Nir Friedman. 1997. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the ICML*. Douglas H. Fisher (Ed.). 125–133.

[56] Nir Friedman, Kevin P. Murphy, and Stuart Russell. 1998. Learning the structure of dynamic probabilistic networks. In *Proceedings of the Uncertainty in Artificial Intelligence*. 139–147.

[57] Tian Gao, Debarun Bhattacharjya, Elliot Nelson, Miao Liu, and Yue Yu. 2022. IDYNO: Learning nonparametric DAGs from interventional dynamic data. In *Proceedings of the International Conference on Machine Learning*. 6988–7001.

[58] Wei Gao and Haizhong Yang. 2022. Time-varying group lasso granger causality graph for high dimensional dynamic system. *Pattern Recognition* 130 (2022), 108789.

[59] Philipp Geiger, Kun Zhang, Bernhard Schölkopf, Mingming Gong, and Dominik Janzing. 2015. Causal inference by identification of vector autoregressive processes with hidden components. In *Proceedings of the International Conference on Machine Learning*. 1917–1925.

[60] Andreas Gerhardus and Jakob Runge. 2020. High-recall causal discovery for autocorrelated time series with latent confounders. In *Proceedings of the Advances in Neural Information Processing Systems*.

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

[62] Chang Gong, Di Yao, Jin Wang, Wenbin Li, Lanting Fang, Yongtao Xie, Kaiyu Feng, Peng Han, and Jingping Bi. 2024. PORCA: Root cause analysis with partially observed data. arXiv:2407.05869. Retrieved from https://arxiv.org/abs/2407.05869

[63] Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, Jingping Bi, Lun Du, and Jin Wang. 2023. Causal discovery from temporal data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5803–5804.

[64] Chang Gong, Di Yao, Lei Zhang, Sheng Chen, Wenbin Li, Yueyang Su, and Jingping Bi. 2024. CausalMMM: Learning causal structure for marketing mix modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. ACM, 238–246.

[65] Mingming Gong, Kun Zhang, Bernhard Schölkopf, Dacheng Tao, and Philipp Geiger. 2015. Discovering temporal causal relations from subsampled data. In *Proceedings of the International Conference on Machine Learning*. Francis R. Bach and David M. Blei (Eds.), 1898–1906.

[66] Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. 2023. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

[67] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 37, 3 (1969), 424–438.

[68] Marco Grzegorczyk. 2024. Being bayesian about learning bayesian networks from ordinal data. *International Journal of Approximate Reasoning* 170 (2024), 109205.

[69] Marco Grzegorczyk and Dirk Husmeier. 2011. Non-homogeneous dynamic bayesian networks for continuous data. *Machine Learning* 83, 3 (2011), 355–419.

[70] Marco Grzegorczyk and Dirk Husmeier. 2013. Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical bayesian models. *Machine Learning* 91, 1 (2013), 105–154.

[71] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2021. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys* 53, 4 (2021), 75:1–75:37.

[72] Hermanni Hälvä and Aapo Hyvärinen. 2020. Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. 939–948.

[73] Peng Han, Jin Wang, Di Yao, Shuo Shang, and Xiangliang Zhang. 2021. A graph-based approach for trajectory similarity computation in spatial networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 556–564.

[74] Niels Hansen and Alexander Sokol. 2014. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability* 19 (2014), 1–24.

[75] Uzma Hasan, Emam Hossain, and Md. Osman Gani. 2023. A survey on causal discovery methods for I.I.D. and time series data. *Trans. Mach. Learn. Res.* 2023 (2023).

[76] Alan G. Hawkes. 1971. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 33, 3 (1971), 438–443.

[77] Shan He, Song Yang, and Dake Chen. 2023. Modeling and prediction of large-scale climate variability by inferring causal structure. *Geophysical Research Letters* 50, 16 (2023), e2023GL104291.

[78] Christina Heinze-Deml, Marloes H. Maathuis, and Nicolai Meinshausen. 2018. Causal structure learning. *Annual Review of Statistics and Its Application* 5 (2018), 371–391.

[79] Miguel A. Hernán and James M. Robins. 2010. Causal inference.

[80] Katerina Hlavackova-Schindler, Kejsi Hoxhallari, Luis Caumel Morales, Irene Schicker, and Claudia Plant. 2024. Causal discovery among wind-related variables in a wind farm under extreme wind speed scenarios: Comparison of results using granger causality and interactive k-means clustering. In *Proceedings of the Many Shades of Causality Analysis in Earth Sciences: Methods, Challenges and Applications*.

[81] Katerina Hlaváčková-Schindler, Anna Melnykova, and Irene Tubikanec. 2024. Granger causal inference in multivariate hawkes processes by minimum message length. *Journal of Machine Learning Research* 25, 133 (2024), 1–26.

[82] Katerina Hlavácková-Schindler and Claudia Plant. 2020. Heterogeneous graphical granger causality by minimum message length. *Entropy* 22, 12 (2020), 1400.

[83] Biwei Huang, Kun Zhang, and Bernhard Schölkopf. 2015. Identification of time-dependent causal model: A gaussian process treatment. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. 3561–3568.

[84] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D. Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. 2020. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research* 21 (2020), 89:1–89:53.

[85] Hao Huang, Chenxiao Xu, Shinjae Yoo, Weizhong Yan, Tianyi Wang, and Feng Xue. 2020. Imbalanced time series classification for flight data analyzing with nonlinear granger causality learning. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. 2533–2540.

[86] Aapo Hyvärinen, Shohei Shimizu, and Patrik O. Hoyer. 2008. Causal modelling combining instantaneous and lagged effects: An identifiable model based on non-gaussianity. In *Proceedings of the 25th International Conference on Machine Learning*. 424–431.

[87] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. 2010. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research* 11 (2010), 1709–1731.

[88] Tsuyoshi Idé, Georgios Kollias, Dzung T. Phan, and Naoki Abe. 2021. Cardinality-regularized hawkes-granger model. In *Proceedings of the Advances in Neural Information Processing Systems*. 2682–2694.

[89] Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. 2022. Root cause analysis of failures in microservices through causal discovery. In *Proceedings of the Advances in Neural Information Processing Systems*.

[90] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery* 33, 4 (2019), 917–963.

[91] Amirkasra Jalaldoust, Katerina Hlavácková-Schindler, and Claudia Plant. 2022. Causal discovery in hawkes processes by minimum description length. 36, 6 (2022), 6978–6987.

[92] Junzhong Ji, Zuozhen Zhang, Lu Han, and Jinduo Liu. 2024. MetaCAE: Causal autoencoder with meta-knowledge transfer for brain effective connectivity estimation. *Computers in Biology and Medicine* 170 (2024), 107940.

[93] Zhuochen Jin, Shunan Guo, Nan Chen, Daniel Weiskopf, David Gotz, and Nan Cao. 2021. Visual causality analysis of event sequence data. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1343–1352.

[94] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation?. In *Proceedings of the International Conference on Learning Representations*.

[95] Alistair EW Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (2016), 1–9.

[96] Marcus Kaiser and Maksim Sipos. 2022. Unsuitability of NOTEARS for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters* 54, 3 (2022), 1587–1595.

[97] Soufiane Karmouche, Evgenia Galytska, Jakob Runge, Gerald A. Meehl, Adam S. Phillips, Katja Weigel, and Veronika Eyring. 2023. Regime-oriented causal model evaluation of atlantic–pacific teleconnections in CMIP6. *Earth System Dynamics* 14, 2 (2023), 309–344.

[98] Mehmet Kayaalp and Gregory F. Cooper. 2002. A bayesian network scoring metric that is based on globally uniform parameter priors. In *Proceedings of the Uncertainty in Artificial Intelligence*. 251–258.

[99] Saurabh Khanna and Vincent Y. F. Tan. 2020. Economy statistical recurrent units for inferring nonlinear granger causality. In *Proceedings of the International Conference on Learning Representations*.

[100] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv* (2023).

[101] Jong-Min Kim, Namgil Lee, and Sun Young Hwang. 2020. A copula nonlinear granger causality. *Economic Modelling* 88 (2020), 420–430.

[102] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N. Brown. 2011. A granger causality measure for point process models of ensemble neural spiking activity. *PLOS Computational Biology* 7, 3 (2011), 1–13.

[103] Neville Kenneth Kitson, Anthony C. Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. 2023. A survey of bayesian network structure learning. *Artificial Intelligence Review* 56, 8 (2023), 8721–8814.

[104] Samantha Kleinberg. 2013. *Causality, Probability, and Time*. Cambridge University Press.

[105] Samantha Kleinberg and Bud Mishra. 2009. The temporal logic of causal structures. In *Proceedings of the Uncertainty in Artificial Intelligence*. 303–312.

[106] Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

[107] Markku Lanne, Mika Meitz, and Pentti Saikkonen. 2017. Identification and estimation of non-gaussian structural vector autoregressions. *Journal of Econometrics* 196, 2 (2017), 288–304.

[108] Steffen L Lauritzen. 1996. *Graphical Models*. Vol. 17. Clarendon Press.

[109] Hongming Li, Shujian Yu, and José C. Príncipe. 2023. Causal recurrent variational autoencoder for medical time series generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8562–8570.

[110] Peiwen Li, Yuan Meng, Xin Wang, Fang Shen, Yue Li, Jialong Wang, and Wenwu Zhu. 2023. Causal discovery in temporal domain from interventional data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. ACM, 4074–4078.

[111] Sha Li, Xiaofeng Gao, Weiming Bao, and Guihai Chen. 2017. FM-hawkes: A hawkes process based approach for modeling online activity correlations. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1119–1128.

[112] Wenbin Li, Di Yao, Chang Gong, Xiaokai Chu, Quanliang Jing, Xiaolei Zhou, Yuxuan Zhang, Yunxia Fan, and Jingping Bi. 2024. Causaltad: Causal implicit generative model for debiased online trajectory anomaly detection. In *Proceedings of the 2024 IEEE 40th International Conference on Data Engineering*. IEEE, 4477–4490.

[113] Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. 2024. STBench: Assessing the ability of large language models in spatio-temporal analysis. arXiv:2406.19065. Retrieved from https://arxiv.org/abs/2406.19065

[114] Cheng-Ming Lin, Ching Chang, Wei-Yao Wang, Kuang-Da Wang, and Wen-Chih Peng. 2024. Root cause analysis in microservice using neural granger causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 206–213.

[115] Dominik Linzner and Heinz Koeppl. 2021. Active learning of continuous-time bayesian networks through interventions. In *Proceedings of the International Conference on Machine Learning*. 6692–6701.

[116] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. 2022. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the International Conference on Machine Learning*. 13557–13603.

[117] Manxia Liu, Fabio Stella, Arjen Hommersom, Peter J. F. Lucas, Lonneke Boer, and Erik Bischoff. 2019. A comparison between discrete and continuous time bayesian networks in learning from clinical time series data with irregularity. *Artif. Intell. Medicine* 95 (2019), 104–117.

[118] Shuo Liu, Di Yao, Lanting Fang, Zhetao Li, Wenbin Li, Kaiyu Feng, XiaoWen Ji, and Jingping Bi. 2024. AnomalyLLM: Few-shot anomaly edge detection for dynamic graphs using large language models. arXiv:2405.07626. Retrieved from https://arxiv.org/abs/2405.07626

[119] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. In *Proceedings of the Socially Responsible Language Modelling Research.*

[120] Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling.*

[121] Edward N Lorenz. 1996. Predictability: A problem partly solved. In *Proc. Seminar on Predictability*, Vol. 1. Reading.

[122] Sindy Löwe, David Madras, Richard Z. Shilling, and Max Welling. 2022. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Proceedings of the Conference on Causal Learning and Reasoning.* 509–525.

[123] Dixin Luo, Hongteng Xu, Hongyuan Zha, Jun Du, Rong Xie, Xiaokang Yang, and Wenjun Zhang. 2014. You are what you watch and when you watch: Inferring household structures from IPTV viewing data. *IEEE Transactions on Broadcasting* 60, 1 (2014), 61–72.

[124] Huishi Luo, Fuzhen Zhuang, Ruobing Xie, Hengshu Zhu, Deqing Wang, Zhulin An, and Yongjun Xu. 2024. A survey on causal inference for recommendation. *The Innovation* 5, 2 (2024).

[125] Daniel Malinsky and Peter Spirtes. 2018. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of the 2018 ACM SIGKDD Workshop on Causal Discovery.* 23–47.

[126] Ricards Marcinkevics and Julia E. Vogt. 2021. Interpretable models for granger causality using self-explaining neural networks. In *Proceedings of the International Conference on Learning Representations.*

[127] Laila Melkas, Rafael Savvides, Suyog Chandramouli, Jarmo Mäkelä, Tuomo Nieminen, Ivan Mammarella, and Kai Puolamäki. 2021. Interactive causal structure discovery in earth system sciences. In *Proceedings of the The KDD'21 Workshop on Causal Discovery.* 3–25.

[128] Giovanni Menegozzo, Diego Dall'Alba, and Paolo Fiorini. 2022. CIPCaD-Bench: Continuous industrial process datasets for benchmarking causal discovery methods. In *Proceedings of the 2022 IEEE 18th International Conference on Automation Science and Engineering.* IEEE, 2124–2131.

[129] Yuan Meng, Shenglin Zhang, Yongqian Sun, Ruru Zhang, Zhilong Hu, Yiyin Zhang, Chenyang Jia, Zhaogang Wang, and Dan Pei. 2020. Localizing failure root causes in a microservice through causality inference. In *Proceedings of the 2020 IEEE/ACM 28th International Symposium on Quality of Service.* 1–10.

[130] Montassar Ben Messaoud, Philippe Leray, and Nahla Ben Amor. 2009. Integrating ontological knowledge for iterative causal discovery and visualization. In *Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty.* 168–179.

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

[132] Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. 2018. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the UAI.* 350–360.

[133] Alessandro Montalto, Sebastiano Stramaglia, Luca Faes, Giovanni Tessitore, Roberto Prevete, and Daniele Marinazzo. 2015. Neural networks with non-uniform embedding and explicit validation phase to assess granger causality. *Neural Networks* 71 (2015), 159–171.

[134] 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. *Knowledge and Information Systems* 63, 12 (2021), 3041–3085.

[135] Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction* 1, 1 (2019), 312–340.

[136] Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, Simon Lacoste-Julien, and Kun Zhang. 2022. On the convergence of continuous constrained optimization for structure learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics.* 8176–8198.

[137] Uri Nodelman, Christian R. Shelton, and Daphne Koller. 2003. Learning continuous time bayesian networks. In *Proceedings of the Uncertainty in Artificial Intelligence.* 451–458.

[138] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery* 12, 2 (2022), e1449.

[139] Rodney T. O'Donnell, Ann E. Nicholson, B. Han, Kevin B. Korb, M. J. Alam, and Lucas R. Hope. 2006. Causal discovery with prior information. In *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19.* 1162–1167.

[140] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. 2020. DYNOTEARS: Structure learning from time-series data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics.* 1595–1605.

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

[142] Judea Pearl and Dana Mackenzie. 2018. *The Book of why: The New Science of Cause and Effect*. Basic books.

[143] Jonas Peters, Stefan Bauer, and Niklas Pfister. 2022. Causal models for dynamical systems. In *Proceedings of the Probabilistic and Causal Inference: The Works of Judea Pearl*. 671–690.

[144] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal inference on time series using restricted structural equation models. In *Proceedings of the NeurIPS*. 154–162.

[145] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

[146] Anne Helby Petersen, Joseph Ramsey, Claus Thorn Ekstrøm, and Peter Spirtes. 2023. Causal discovery for observational sciences using supervised machine learning. *Journal of Data Science* 21, 2 (2023), 255–280.

[147] Bernd Pompe and Jakob Runge. 2011. Momentary information transfer as a coupling measure of time series. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 83, 5 (2011), 051122.

[148] Robert J. Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. 2010. Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PloS One* 5, 2 (2010), e9202.

[149] Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. 2021. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. In *Proceedings of the Advances in Neural Information Processing Systems*. 27772–27784.

[150] Weijie Ren, Baisong Li, and Min Han. 2020. A novel granger causality method based on HSIC-lasso for revealing nonlinear relationship between multivariate time series. *Physica A: Statistical Mechanics and its Applications* 541 (2020), 123245.

[151] Joshua W. Robinson and Alexander J. Hartemink. 2010. Learning non-stationary dynamic bayesian networks. *Journal of Machine Learning Research* 11 (2010), 3647–3680.

[152] Raanan Y. Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. 2021. Iterative causal discovery in the possible presence of latent confounders and selection bias. In *Proceedings of the Advances in Neural Information Processing Systems*. 2454–2465.

[153] Raanan Y. Yehezkel Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. 2023. From temporal to contemporaneous iterative causal discovery in the presence of latent confounders. In *Proceedings of the International Conference on Machine Learning*. PMLR, 39939–39950.

[154] Paul K. Rubenstein, Stephan Bongers, Joris M. Mooij, and Bernhard Schölkopf. 2018. From deterministic ODEs to dynamic structural causal models. In *Proceedings of the Uncertainty in Artificial Intelligence*. 114–123.

[155] Jakob Runge. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. 1388–1397.

[156] Jakob Runge, Jobst Heitzig, Norbert Marwan, and Jürgen Kurths. 2012. Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy. *Physical Review E* 86, 6 (2012), 061121.

[157] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances* 5, 11 (2019), eaau4996.

[158] Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. 2020. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30, 11 (2020), 113115.

[159] Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. 2019. A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence* 8, 4 (2019), 425–439.

[160] Thomas Schreiber. 2000. Measuring information transfer. *Physical Review Letters* 85, 2 (2000), 461.

[161] Patrick Schwab, Djordje Miladinovic, and Walter Karlen. 2019. Granger-causal attentive mixtures of experts: Learning important features with neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4846–4853.

[162] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. 2019. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning* 115 (2019), 235–253.

[163] Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Guangyin Jin, Xin Cao, Gao Cong, et al. 2024. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *TKDE* (2024), 1–14.

[164] Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. 2022. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 1567–1577.

[165] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti J. Kerminen. 2006. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7 (2006), 2003–2030.

[166] Ali Shojaie and Emily B Fox. 2022. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application* 9, 1 (2022), 289–319.

[167] Xiao Shou, Debarun Bhattacharjya, Tian Gao, Dharmashankar Subramanian, Oktie Hassanzadeh, and Kristin P. Bennett. 2023. Probabilistic attention-to-influence neural models for event sequences. In *Proceedings of the International Conference on Machine Learning*. PMLR, 31657–31674.

[168] Xiao Shou, Tian Gao, Dharmashankar Subramanian, Debarun Bhattacharjya, and Kristin P. Bennett. 2023. Influence-aware attention for multivariate temporal point processes. In *Proceedings of the Conference on Causal Learning and Reasoning*. PMLR, 499–517.

[169] Nitin K. Singh and David M. Borrok. 2019. A granger causality analysis of groundwater patterns over a half-century. *Scientific Reports* 9, 1 (2019), 12828.

[170] Stephen M. Smith, Karla L. Miller, Gholamreza Salimi Khorshidi, Matthew A. Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark William Woolrich. 2011. Network modelling methods for FMRI. *NeuroImage* 54, 2 (2011), 875–891.

[171] Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search, Second Edition.*

[172] Chandler Squires and Caroline Uhler. 2023. Causal structure learning: A combinatorial perspective. *Foundations of Computational Mathematics* 23, 5 (2023), 1781–1815.

[173] James V. Stone. 2004. Independent component analysis: A tutorial introduction. (2004).

[174] Liessman Sturlaugson and John W. Sheppard. 2014. Inference complexity in continuous time bayesian networks. In *Proceedings of the UAI*. 772–779.

[175] George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. 2012. Detecting causality in complex ecosystems. *Science* 338, 6106 (2012), 496–500.

[176] Jie Sun and Erik M. Bollt. 2014. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena* 267 (2014), 49–57.

[177] Jie Sun, Dane Taylor, and Erik M. Bollt. 2015. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems* 14, 1 (2015), 73–106.

[178] Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. 2023. NTS-NOTEARS: Learning nonparametric temporal DAGs with time-series data and prior knowledge. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 206. 1942–1964.

[179] Floris Takens. 1981. Detecting strange attractors in turbulence. In *Proceedings of the Dynamical Systems and Turbulence, Warwick 1980*. 366–381.

[180] Alex Tank, Ian Covert, Nicholas J. Foti, Ali Shojaie, and Emily B. Fox. 2022. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2022), 4267–4279.

[181] Alex Tank, Emily B Fox, and Ali Shojaie. 2019. Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series. *Biometrika* 106, 2 (2019), 433–452.

[182] Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. 2022. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science* 1 (2022), e12.

[183] Gherardo Varando, Miguel-Angel Fernández-Torres, and Gustau Camps-Valls. 2021. Learning granger causal feature representations. In *Proceedings of the ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.

[184] Simone Villa and Fabio Stella. 2016. Learning continuous time bayesian networks in non-stationary domains. *Journal of Artificial Intelligence Research* 57 (2016), 1–37.

[185] Carlos Villa-Blanco, Alessandro Bregoli, Concha Bielza, Pedro Larrañaga, and Fabio Stella. 2023. Constraint-based and hybrid structure learning of multidimensional continuous-time bayesian network classifiers. *International Journal of Approximate Reasoning* 159 (2023), 108945.

[186] Mark Voortman, Denver Dash, and Marek J. Druzdzel. 2010. Learning why things change: The difference-based causality learner. In *Proceedings of the Uncertainty in Artificial Intelligence*. 641–650.

[187] Matthew J. Vowels, Necati Cihan Camgöz, and Richard Bowden. 2023. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys* 55, 4 (2023), 82:1–82:36.

[188] Dongjie Wang, Zhengzhang Chen, Yanjie Fu, Yanchi Liu, and Haifeng Chen. 2023. Incremental causal graph learning for online root cause analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2269–2278.

[189] Dongjie Wang, Zhengzhang Chen, Jingchao Ni, Liang Tong, Zheng Wang, Yanjie Fu, and Haifeng Chen. 2023. Interdependent causal networks for root cause localization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 5051–5060.

[190] Yueming Wang, Kang Lin, Yu Qi, Qi Lian, Shaozhe Feng, Zhaohui Wu, and Gang Pan. 2018. Estimating brain connectivity with varying-length time lags using a recurrent neural network. *IEEE Transactions on Biomedical Engineering* 65, 9 (2018), 1953–1963.

[191] Song Wei, Yao Xie, Christopher S. Josef, and Rishikesan Kamaleswaran. 2023. Granger causal chain discovery for sepsis-associated derangements via continuous-time hawkes processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2536–2546.

[192] Tianhao Wu, Xingyu Wu, Xin Wang, Shikang Liu, and Huanhuan Chen. 2022. Nonlinear causal discovery in time series. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. 4575–4579.

[193] Chenxiao Xu, Hao Huang, and Shinjae Yoo. 2019. Scalable causal graph learning through a deep neural network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1853–1862.

[194] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. 2016. Learning granger causality for hawkes processes. In *Proceedings of the International Conference on Machine Learning*. 1717–1726.

[195] Yongjun Xu, Fei Wang, Zhulin An, Qi Wang, and Zhao Zhang. 2023. Artificial intelligence for science—bridging data to wisdom. *The Innovation* 4, 6 (2023), 100525.

[196] Di Yao, Gao Cong, Chao Zhang, Xuying Meng, Rongchang Duan, and Jingping Bi. 2020. A linear time approach to computing time series similarity based on deep metric learning. *TKDE* 34, 10 (2020), 4554–4571.

[197] Di Yao, Chang Gong, Lei Zhang, Sheng Chen, and Jingping Bi. 2022. CausalMTA: Eliminating the user confounding bias for causal multi-touch attribution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 4342–4352.

[198] Di Yao, Haonan Hu, Lun Du, Gao Cong, Shi Han, and Jingping Bi. 2022. TrajGAT: A graph-based long-term dependency modeling approach for trajectory similarity computation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2275–2285.

[199] Di Yao, Jin Wang, Wenjie Chen, Fangda Guo, Peng Han, and Jingping Bi. 2024. Deep dirichlet process mixture model for non-parametric trajectory clustering. In *Proceedings of the 2024 IEEE 40th International Conference on Data Engineering*. IEEE, 4449–4462.

[200] Di Yao, Chao Zhang, Jianhui Huang, and Jingping Bi. 2017. Serm: A recurrent model for next location prediction in semantic trajectories. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2411–2414.

[201] Di Yao, Chao Zhang, Zhihua Zhu, Jianhui Huang, and Jingping Bi. 2017. Trajectory clustering via deep representation learning. In *Proceedings of the 2017 International Joint Conference on Neural Networks*. IEEE, 3880–3887.

[202] Weiran Yao, Guangyi Chen, and Kun Zhang. 2022. Temporally disentangled representation learning. In *Proceedings of the Advances in Neural Information Processing Systems*.

[203] Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, and Yongjun Xu. 2024. GinAR: An end-to-end multivariate time series forecasting model suitable for variable missing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 3989–4000.

[204] Chengqing Yu, Fei Wang, Zezhi Shao, Tao Sun, Lin Wu, and Yongjun Xu. 2023. DSformer: A double sampling transformer for multivariate time series long-term prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. ACM, 3062–3072.

[205] Jing Yu, V. Anne Smith, Paul P. Wang, Alexander J. Hartemink, and Erich D. Jarvis. 2004. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 20, 18 (2004), 3594–3603.

[206] Xiufan Yu, Karthikeyan Shanmugam, Debarun Bhattacharjya, Tian Gao, Dharmashankar Subramanian, and Lingzhou Xue. 2020. Hawkesian graphical event models. In *Proceedings of the International Conference on Probabilistic Graphical Models*. 569–580.

[207] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning* 151 (2022), 101–129.

[208] Cheng Zhang, Dominik Janzing, Mihaela van der Schaar, Francesco Locatello, and Peter Spirtes. 2023. Causality in the time of LLMs: Round table discussion results of CLeaR 2023. *Proceedings of Machine Learning Research* (2023).

[209] Jiji Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* 172, 16-17 (2008), 1873–1896.

[210] Jiji Zhang and Peter Spirtes. 2008. Detection of unfaithfulness and robust causal inference. *Minds Mach.* 18, 2 (2008), 239–271.

[211] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the IJCAI*. 1347–1353.

[212] Wei Zhang, Thomas Kobber Panum, Somesh Jha, Prasad Chalasani, and David Page. 2020. CAUSE: Learning granger causality from event sequences using attribution methods. In *Proceedings of the International Conference on Machine Learning*. 11235–11245.

[213] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. DAGs with NO TEARS: Continuous optimization for structure learning. In *Proceedings of the Advances in Neural Information Processing Systems*. 9492–9503.

[214] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning social infectivity in sparse low-rank networks using multidimensional hawkes processes. In *Proceedings of the Artificial Intelligence and Statistics*. 641–649.

[215] Hua Zhu, Hong Huang, Kehan Yin, Zejun Fan, Hai Jin, and Bang Liu. 2024. CausalNET: Unveiling causal structures on event sequences by topology-informed causal attention. In *Proceedings of the IJCAI*. 7144–7152.

Received 26 February 2024; revised 14 September 2024; accepted 3 November 2024