A Survey on Causal Discovery with Incomplete Time-Series Data*

Xuanzhi Chen, Wei Chen and Ruichu Cai

Data mining and Information Retrieval Lab
School of Computer Science, Guangdong University of Technology
Guangzhou, China
xuanzhichen.42@gmail.com/xuanzhichen.github.io

Abstract

With the rapid growth of massive time-series data, inferring temporal Bayesian structures based on causation from data — Temporal Causal Discovery (TCD) — has become an important and challenging task in recent years. It holds increasing scientific significance and commercial value for computationally uncovering structural knowledge and/or generative mechanisms behind the data. **Although existing reviews have systematically introduced TCD methods based on observational data, they have not fully considered the issues of incompleteness caused by hidden variables (latent confounders) or data missing.** In this review, we focus on the least research progress in the task of TCD with incomplete data, summarizing the philosophy and paradigms reflected in current research methods. To this end, we elaborate on causality algorithms applicable to incomplete time-series data within the categories of **four theoretical frameworks**: Conditional Independence Tests (CIT), Structural Causal Models (SCM), Score Functions (SF), and Granger Causality (GC). We further introduce how TCD algorithms address challenges in **two real-world applications**, namely non-uniform sampling and non-stationarity. Additionally, we list common case studies and typical evaluation metrics related to TCD with incomplete data. Finally, in discussing future research direction, we call on that, beyond parallel advancements upon classic methods and excessive reliance on function fitting, **it will be essential for innovative approaches capable of handling "the Time Dimension" in a targeted and transparent way** — the core perspective in this review that warrants further exploration in causation. Materials such as brief presentation slides for this paper are available on the author's website: xuanzhichen.github.io (Work/Paper/2023).

Keywords

Time-Series Causal Discovery, Bayes-Nets Structural Learning, Hidden Variables, Missing Data

1. Introduction

In recent years, high-volume and high-dimensional data have been continuously developing. Time series data, as an ordered sequence of real values collected over time, often carries real-world information. Over the past few decades, numerous time series analysis methods based on various tasks such as classification [37][41], clustering [1][40], and forecasting [69][70] have emerged. Among these, temporal causal discovery (TCD) based on observational data—identifying causal-and-effect between different time series without relying on intervention—is also a notable task pertaining to explainable data.

Admittedly, the gold standard for inferring causal relationships between different variables is widely accepted as Randomized Control Trials (RCTs) [30]. However, due to inherent drawbacks such as high costs, long durations, and ethical concerns associated with RCTs, a significant number of excellent causal discovery methods for non-experimental data have been proposed over the past decade [68][24]. The results of these methods are often described in the form of Directed Acyclic Graphs (DAGs), where directed edges between variables represent the asymmetric relation (causation) rather then correlations.

* *A Publication Template by CEUR-WS (https://ceur-ws.org/) in a Conference Proceedings.*

This research review has been a finished manuscript that should have braced for submission in 2023, but collaboration between Xuanzhi Chen and other authors came to an early cessation, due to the personal inconvenience of Xuanzhi Chen.

***Personal Academic Statement**: Xuanzhi Chen renounce the use of any AI-generative content throughout in this paper, though it is predictable that, in the future, AI would be better at onerous and comprehensive work like research review than humans.*

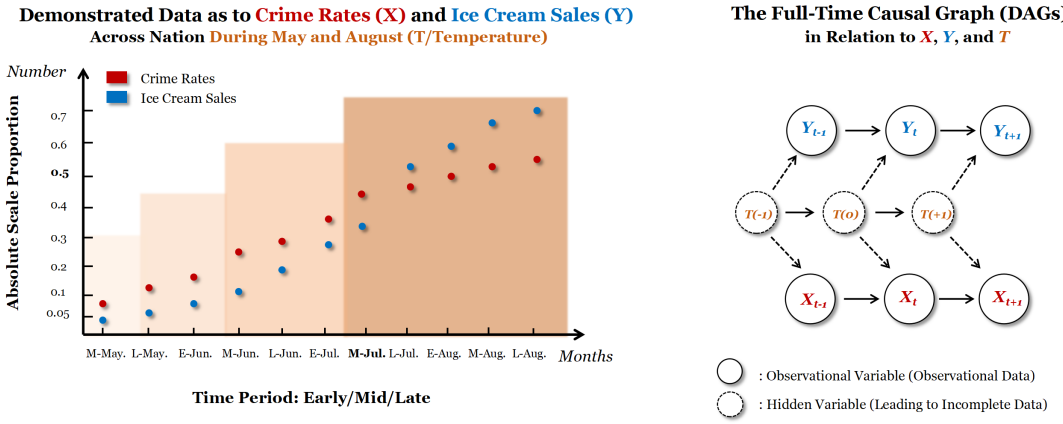

Figure 1: A Time-Series-Based System lacking Consideration of Causality (Thought Experiments).

When it comes to the context of time series, in particular considering the autocorrelation among variables, nevertheless, causal graphs are typically aggregated or expanded along the time axis based on time windows, which further categorizing them into: Summary Causal Graphs, Window Causal Graphs, and Full-Time Causal Graphs. Different types of temporal causal graphs serve as the objectives for TCD in numerous systems of natural sciences or human society, such as climate science [62], efficacy assessment [8], and economic markets [32], adding complexities as to dynamic systems analysis.

On the other hand, time-series-based systems that lack consideration of causality may mislead people to draw incorrect or even absurd conclusions. **For example (Figure 1), the sales of ice cream in urban areas during the summer and the rise in local crime rates are clearly two time series variables that intuitively seem unrelated (non-causal), yet they may exhibit a high statistical correlation.** In fact, merely considering the time series of ice cream sales and crime rates as observational data is INCOMPLETE. A reasonable explanation is that the unobserved data related to urban temperatures leads to this biased result; baking heat during urban hot summers simultaneously stimulate increases in both ice cream sales and potential crime rates. In order words, the factor of urban temperature acts as a hidden variable (latent confounder), causing a spurious correlation between the time series of ice cream sales and crime rates in statistical analysis.

So far, several reviews [48][4][25][31] have comprehensively summarized the progress in TCD based on observational data from uniquely different perspectives, including causal-effect analysis [48], practical applications of various causality methods [4], data type of event-sequence[31], and the integration of non-temporal and temporal causal discovery issues [31]. However, topics regarding TCD based on incomplete data — observational data with hidden variables or missing samples — are still insufficient. Therefore, we in the paper are meant to investigate the least research progress in relation to this kind of topic, as it is shown in Table 1.

Table 1
The Least Research Progress in the Task of TCD (Temporal Causal Discovery) with Incomplete Data

Causality Algorithms	Four Theoretical Frameworks				Two Real-World Applications	
	CIT-Based	SCF-Based	SF-Based	GC-Based	Non-Uniform Sampling	Non-Stationarity
Classic CD/TCD	PC[61] FCI[61] PCMCI[55] PCMCI-Plus[56]	LiNGAM[58][59] ANMs[33] VAR-LiNGAM[36]	GES[12] Notears[76] DyNotears[50]	GC[28] C-GC[22][9][6] Lasso-GC[60] Kernel-GC[44] TE-GC[5][64][65]	NG-EM[27][26] NG-MF[27][26] CUTS[11]	GP-Model[34] BackShift[54] CD-NOD[35]
TCD with Incomplete Data	LPCMCI[21] ts-FCI[16] SVAR-FCI[43] Tiered Background Knowledge[2] DMAGs/DPAGs[20] SyPI[45]	ANLTSM[13] VAR with Hidden Components[19] TiMINo[52]	SEM[18] SVAR-GFCI[43] ANLTSM-FCI[13]	Partial-GC[29] Eliminating Confounding GC[3] PDC-GC[15] TCDF[49] V-Granger[47] Deep recurrent GC[74]		

Technically, the challenges of incomplete data for TCD arise from theoretical interference due to temporal latent confounders and practical issues from irregular data, stemming from sampling frequency limitations or non-stationarity from environmental shifts. To cope with these challenges, most existing TCD methods can be theoretically categorized into the following four framework bases: conditional independence tests, structural causal models, score functions, and Granger Causality, along with specific strategies in real-world applications. Other notable TCD ideas include discrete logic trees[38][75], differential equation modeling frameworks [42][23], methods based on Takens' theorem [66], and manifold space analysis [63]. Since many of these methods require no unobserved temporal variables or assume prior domain knowledge, we will focus on the four aforementioned frameworks.

Last but not least, we note that current mainstream TCD methods often treat time series as a separate dimensional augmentation from non-temporal data, which aligns with paralle advancements in non-temporal algorithms. Yet fundamentally, their ability to discover causal relations based on incomplete data is the same as that of classical methods. On the other hand, while deep learning techniques such as neural networks, aligning well with the Granger causality theories, can effectively fit unobserved data, causal implications found by these neural networks remain controversial. In conclusion, leveraging temporal information more effectively to improve the performance and interpretability of TCD methods presents both a promising direction and a potential challenge for future research.

2. The Least Research Progress of TCD with Incomplete Data

Commencing with the following section, we will formally elaborate upon the research progress in temporal causal discovery (TCD) in form of **technical reports**, in which **philosophies and paradigms as to the TCD algorithms with incomplete data** will be highlighted between the lines.

**Reminder: Since the discussion with incomplete time series data serves as a relatively small area in fields of causal discovery, we assume readers holds prior knowledge as to non-temporal (classic) and temporal causality approaches; readers may jump to Discussion and Conclusion for general ideas of this paper.*

2.1. Causal Discovery Algorithms with Latent Variables over Time-Series Data

2.1.1. Methods Based on Conditional Independence Tests

The methodology leveraging conditional independence tests (CIT), known as the constraint-based approach, extends Bayesian structural learning methods within causal significance constraints. In the early 1980s, researchers like Glymour and Spirtes developed efficient causal discovery algorithms that utilize statistical patterns of (conditional) independence and incorporate completeness through philosophical logic rules.

Given the assumption of Causal Sufficiency [61], fundamental approaches encompass the SGS (Spirtes-Glymour-Scheines) algorithm and the PC (Peter-Clark) algorithm [61]. Equation (1) represents the CIT for any extant variable pair x_i, x_j conditioned upon (every subset consisting of) variables other than x_i, x_j within the observed variable set V.

$$CI\Big(x_i,\ x_j \mid subset\Big(\ \mathbf{x}^{V\setminus\{x_i,\ x_j\}}\Big)\Big)\,. \tag{1}$$

Unlike classic CIT, momentary conditional independence tests (MCI test), shown in Equation (2), are designed to bolster up the discovery rate of the causal relations (e.g. $x^j \to x^i$) whose statistical variation is impacted by autocorrelation (e.g. time lag $p = 1, 2, 3...$) in the context of time series. We will back to this comparation soon as we introduce the PCMCI algorithms[21].

$$MCI\Big(x_t^i,\ x_{t-p}^j \mid \mathbf{pa}_t^i \setminus \big\{x_{t-p}^j\big\},\ \mathbf{pa}_{t-p}^j\Big)\,. \tag{2}$$

Constraint-based methods rely on the Markov and Faithfulness assumptions, decomposing the learning process into two stages: skeleton learning and direction orientation based on the V-structure [51]. Since multiple causal structures can share identical independence patterns, leading to partial orientation, the goal is to construct Markov equivalent classes (MECs) over partially directed acyclic graphs (PDAGs), also known as completed partially directed acyclic graphs (CPDAGs). Concretely, commencing with a complete graph, execution of CIT eliminates redundant edges between pairwise variables, yielding a causal skeleton. The algorithms then orient the edge direction mainly in light of the V-structure provided by the condition (separation) set, ultimately leading to CPDAGs.

In terms of systems that involve incomplete data and fail to satisfy the Causal Sufficiency assumption, the mainstream approach, represented by the FCI (Fast Causal Inference) algorithm[61], is adopted and proven to be theoretically correct, sound, and complete. The FCI algorithm is an extension of the PC algorithm, which introduces the concept of maximal ancestral graphs (MAGs) and possible d-separation sets (Possible-Dsep sets) to aid in respectively representing the causal graphs and testing conditional independence in the presence of latent confounders. On one hand, similar to the PC algorithm, the FCI algorithm also employs the V-structure and the logical rule to determine causal direction over MAGs, leading to the search of partially ancestral graphs (PAGs) that are analogous to PDAGs. On the other hand, to overcome the potential inefficiency of searching across large combinations of the Possible-Dsep sets, related work[14] further proposes RFCI (Really Fast Causal Inference) to avoid traditional extensive search by additionally introducing particular independence tests.

In light of the aforementioned approaches (PC and FCI algorithms), recent advancements in TCD with incomplete data have emerged as extensions of traditional constraint-based methods, subject to the same fundamental assumptions (Markov and Faithfulness assumptions). These notable advancements encompass two primary approaches: (1) The LPCMCI algorithm[21] based on the PC algorithm. (2) The SVAR-FCI algorithm[43] based on the FCI algorithm.

One may notice that current TCD algorithms often derive a window causal graph by discovering causal relations within a certain time-delay range. Yet variables that fall outside this delay range, especially those acting as latent confounders, can propagate their confounding effects through the autocorrelation of time series variables. This necessitates researchers to introduce higher-order Markov assumptions to capture more information, leading to the LPCMCI (Latent PCMCI) algorithm that considers the influence of long-range causal effects through variable parameter settings. The key idea of LPCMCI lies in its ability to enhance the discovery rate of the causal edges affected by the autocorrelation through the momentary conditional independence test (MCI test), which we have illustrated previously in formula (2).

Building on this, LPCMCI[21] further generalizes the application of the MCI test by shrinking or expanding the size of the conditioning set. The LPCMCI strategy to handle latent variables is fundamentally based on an information-theoretic perspective, indicating that if ancestral variables causing long-range causal effects can be incorporated into the CIT as early as possible, the effect size[21] of the conditioning set (in the present of latent variables) can be significantly improved. Formula (3) illustrates how to optimize the conditioning set S to achieve the best effect size:

$$S = \arg\min_{S \subseteq \mathbf{x}^*} I\left(x_t^i ; x_{t-p}^j \mid S \cup S_{def}^{\mathscr{M}}\right). \tag{3}$$

where $\mathbf{x}^* = \mathbf{x}^{V \setminus \{x_t^i,\, x_{t-p}^j\}}$, $S_{def}^{\mathscr{M}} = \{\mathbf{pa}_t^i,\ \mathbf{pa}_{t-p}^j\} \setminus \{x_t^i,\ x_{t-p}^j\}$. $S_{def}^{\mathscr{M}}$ is updated from S and represents the default conditioning set obtained from the current (estimated) maximal ancestral graph (MAGs).

More specifically, as shown in formula (4), the LPCMCI method shows that if the conditioning set considers only ancestral variables, it will improve the upper bound of mutual information for a pairwise variables, avoiding the introduction of spurious associations from latent variables.

$$\min_{S} I\left(x_t^i ; x_{t-p}^j \mid S \cup S_{def}^{\mathcal{M}}\right) > \min_{S'} I\left(x_t^i ; x_{t-p}^j \mid S'\right). \tag{4}$$

To achieve this, LPCMCI defines middle marks and novel constraint rules to represent potential ancestral relationships, allowing LPCMCI to identify the corresponding MAGs and discover potential latent ancestral variables by MCI tests as early as possible in the algorithm's process of removing redundant edges. Unlike traditional constraint-based methods with a sequential order, LPCMCI adopts an iterative updating algorithm framework until the algorithm converges to the optimal PAGs. Meanwhile, because of the use of this dynamically optimized framework, LPCMCI is thus order-independent, sound, and complete, as compared to traditional PC algorithms.

On the other hand, as relatively prior constraint-based TCD methods for handling the latent confounder, the ts-FCI (Time Series FCI) algorithm[16] was posited as an extension to the FCI algorithm. The ts-FCI technique[16], predicated on the stationary assumption and parameter fixing, involves partitions and regulations of time series through sliding windows and time lag. This results in several partitioned variable sets that facilitate the direct implementation of the FCI algorithm over time series data, given assumptions of homogeneity and temporal precedence. Namely, based on the maximum time lag P, the original time series vector $\mathbf{x}_t = \left(x_t^1, ..., x_t^d\right)_{t_0 \leqslant t \leqslant T}$ can be transformed into the (more available) one with a length Pd and a total sample size $T - P$. However, due to the absence of instantaneous causal effect within ts-FCI, SVAR-FCI(Structure Vector Auto Regression FCI)[43] takes into account the interplay between the propagation interval and the sampling interval during the modeling process, reducing the impact from contemporaneous latent confounders. Expanding upon the ts-FCI algorithm, SVAR-FCI enriches the types of temporal latent confounders that can be handled with the realm of constraint-based methods. It is imperative to acknowledge that the ability of ts-FCI and SVAR-FCI to unveil latent variables essentially derives from the FCI algorithm.

From the aforementioned SVAR-FCI algorithm and LPCMCI algorithm, it is evident that methods based on CIT are closely related to the definitions and assumptions of graphical structures. In this regard, some related work considers how to better integrate background knowledge with graphical models in time series, and how the graphical models themselves can exhibit stronger performance in the presence of temporal latent confounders.

According to the Meek criterion, researchers in [2] point out that if background knowledge is available, similar prior information can be integrated into MAGs-based TCD methods in a manner called tiered background knowledge, thereby identifying more previously unrecognized causal relationships affected by (instantaneous) latent confounders. The researchers prove that tiered background knowledge is sound and complete for the ts-FCI and SVAR-FCI algorithms.

Similarly, because direct representation of causal graphs does not require additional assumptions, some work also approaches from the perspective of graphical models [20], indicating that methods based on CIT still have room for improvement in their ability to model latent confounders in temporal scenarios using traditional ancestral graph modeling. Based on subclasses within ancestral graphs, directed maximal ancestral graphs (DMAGs) and directed partial ancestral graphs (DPAGs) are defined[20], allowing these types of graphical representations to convey richer and more accurate information in the presence of temporal latent confounders.

Although methods based on CIT have been expanded and applied in TCD, the existence of the time dimension has also led to more variable combinations as the search space for conditioning sets, which may result in higher computational complexity and lower accuracy in CIT. Recognizing the importance of conditioning variable selection in temporal contexts, related work [45] investigates how latent confounded temporal causal paths cannot be reflected in summary graphs when aggregated, thereby defining the concept of sg-unconfounded causal paths. Consequently, this work proposes the SyPI algorithm, which performs feature selection for conditioning sets based on sg-unconfounded causal paths, narrowing the search space for testing conditioning sets, and alleviating statistical errors that often arise in practical applications of hypothesis testing.

Finally, it is evident that algorithms discussed in this section aim to apply CIT under Markov and faithfulness assumptions to identify causation among multiple variables. By controlling for the autocorrelation of time series variables and the delay range, methods based on CIT can minimize the impact of temporal latent confounding effects as much as possible in temporal situations. Simultaneously, given the rich variety of expressions in graphical models under temporal contexts, these methods have the potential to more accurately capture and embed information regarding temporal latent confounders. However, due to the inherent challenges of Markov equivalence classes (MEC), structural uniqueness is still not guaranteed. The next section will specifically discuss methods based on structural causal models, where researchers typically resort to for a finely grained expression of causal relations.

2.1.2. Methods Based on Structural Causal Models

The core of methods based on structural causal models (SCM) lies in establishing a functional causal model (FCM), which interprets a causal system as a series of special equations — each of them explains the generation of variables as a result of their direct causes and independent noise terms, mapped through an irreversible causal function. The inherent uniqueness of the variable generation mechanism described by the FCM is also known as causal asymmetry, meaning that a variable Y in the system responds only to changes in X, and not vice versa. However, in specific real-world scenarios, if researchers wish to explicitly discover causal relationships only from observational data, and to ensure that the learned causal structure is unique, then additional assumptions must be added to the FCM. Contrary to intuition, the noise terms, or disturbance that cause individual differences within SCM, are actually beneficial for inferring causal directions from observation and maintaining structural uniqueness..

One typical form of causal discovery from observational data based on SCM is the Linear non-Gaussian Acyclic Model (LiNGAM), which relies on the non-Gaussianity of noise. The basic idea of the LiNGAM is that the asymmetry inherent in the assumption of independent non-Gaussian noise mathematically allows for the identification of causal relationships that traditional Linear Gaussian Bayesian networks cannot achieve, enabling causal discovery under the constraints of DAGs. In the LiNGAM, the SCM is represented as follows:

$$\mathbf{x} := B\mathbf{x} + \mathbf{n}, \tag{5}$$

where B and $\mathbf{n}$ represent the lower-triangular causal adjacency matrix (namely the specific form of the FCM) and the corresponding non-Gaussian independent noise vector of the observed variable vector $\mathbf{x}$. The mainstream methods for solving the LiNGAM include two categories. The first category equivalently transforms the LiNGAM into a standard linear Independent Component Analysis (ICA) model, and uses corresponding statistical techniques to solve this linear system [58]. The second category directly resorts to least squares estimation or maximum likelihood estimation methods to search for the most reasonable causal ordering [59].

In addition to non-Gaussianity, another typical form is the Additive Noise Model (ANM) [33], which constrains the FCM by the third derivative of nonlinear function f, along with a broader summary[10]:

$$\mathbf{x} := f(\mathbf{pa}_{\mathbf{x}}) + \mathbf{n}, \tag{6}$$

where $\mathbf{pa}_{\mathbf{x}}$ and $\mathbf{n}$ represent the direct parents of the observed variable vector and the corresponding Gaussian or non-Gaussian independent noise.

Given the successful application of the LiNGAM based on non-Gaussianity in non-temporal scenarios, the VAR-LiNGAM [36] serves as a generalization of the LiNGAM for TCD. The VAR-LiNGAM effectively combines non-Gaussianity with structural vector autoregression (SVAR) models, and proposes a two-stage algorithm that combines AR model estimation and LiNGAM analysis to estimate both delayed and instantaneous causation respectively. In the VAR-LiNGAM, the SCM is represented as a linear non-Gaussian system in the following random process:

$$\mathbf{x}_t := \sum_{p=1}^{P} B_p \mathbf{x}_{t-p} + B_0 \mathbf{x}_t + \mathbf{n}_t, \tag{7}$$

where B_p represents the causal adjacency matrix at delay p and $\mathbf{n}_t$ denotes the non-Gaussian independent noise in the random process. It is important to note that when the delay $p = 0$, B_0 models the instantaneous causal effects based on DAGs, in the sense that B_0 is also represented as a lower-triangular causal adjacency matrix similar to that in LiNGAM. Specifically, in the two-stage solution process of the VAR-LiNGAM, as shown in formula (8), the classical AR model is first used to obtain B'_p as the least-squares fit for the delayed adjacency matrix B_p (for $p > 0$), namely $\mathbf{x}_t := \sum_{p=1}^{P} B'_p \mathbf{x}_{t-p} + \mathbf{n}'_t$:

$$\mathbf{x}_t := \sum_{p=0}^{P} \underbrace{(I - B_0)^{-1} B_p}_{B'_p} \mathbf{x}_{t-p} + \underbrace{(I - B_0)^{-1} \mathbf{n}_t}_{\mathbf{n}'_t}. \tag{8}$$

Furthermore, the residuals model $\hat{\mathbf{n}}'_t = \mathbf{x}_t - \sum_{p=1}^{P} \hat{B}'_p \mathbf{x}_{t-p}$ are analyzed using standard LiNGAM analysis, estimating the instantaneous causal effect matrix B_0 (for $p = 0$) for $\hat{\mathbf{n}}'_t := B_0 \hat{\mathbf{n}}'_t + \mathbf{n}_t$. Based on the estimates of instantaneous causal effects (B_0), the final estimates for delayed causal effects are obtained as $\hat{B}_p = (I - \hat{B}_0)\hat{B}'_p (p > 0)$.

The LiNGAM has other variants in the context of time series. For instance, in the Multi-dimensional LiNGAM [57], data in various dimensions is allowed to exist in the form of tensor, and causal information exposed during the conversion of tensors to matrices is used for structural inference. Furthermore, to make the VAR-LiNGAM applicable to nonlinear systems in time series data, related work [72] proposed the TCD method called NCDH, which incorporates techniques of nonlinear ICA.

Building on the models introduced above, the subsequent sections will focus on the performance and application of TCD methods when faced with incomplete data. The ANLTSM (Additive Non-Linear Time Series Model) proposed in related work [13] first establishes an additive noise model with instantaneous latent confounders based on the VAR model:

$$x_t = B_t x_t + \sum_{p=1}^{P} f(\mathbf{x}_{t-p}) + C_t \mathbf{u}_t + \mathbf{n}_t. \tag{9}$$

Here, $\mathbf{u}_t$ represents the instantaneous latent confounder, $f(\cdot)$ is a smooth multivariate nonlinear function, and $\mathbf{u}_t, \mathbf{n}_t$ are independent Gaussian noise, with distributions $\mathbf{n}_t \sim \mathcal{N}(0, \sigma_1^2)$, $\mathbf{u}_t \sim \mathcal{N}(0, \sigma 2^2)$.

The key idea of the ANLTSM is to efficiently obtain the conditional independence constraint — information that would originally require numerous conditional independence tests (CIT) — through obtaining residuals based on additive regression. The ANLTSM demands the stationary assumption and the Causal Sufficiency assumption, under which the nonparametric estimation of Conditional Expectations for (observational) nonlinear-and-delay time series is asymptotically consistent [13]. This is also why ANLTISM needs to constrain latent confounders only in linear and instantaneous types to ensure the reliability of additive regression. For example, the information obtained from the conditional independence test $CI(x_t^i, x_t^j \mid \mathbf{x}_t^{V\setminus\{i,j\}})$ can be approximated by testing the conditional expectation $\mathbb{E}\left[x_t^i \mid x_t^j,\ \mathbf{x}_t^{V\setminus\{i,j\}},\ \mathbf{x}_{t-p}^{V}\right]$. Furthermore, the independence information involving instantaneous and delayed relationships obtained through two special additive regressions will be input into the FCI algorithm, allowing the FCI algorithm to be applicable for the TCD task with latent variables.

Regarding the modeling of temporal latent confounders, related work [19] simultaneously models the transition matrices of temporal observed variables and latent variables. For example, in formula (10), the VAR-based model shows that when the submatrix $C \neq 0$, it indicates that the current system is influenced by the latent confounder Z (the submatrix B represents the $d \times d$ causal effect matrix):

$$\begin{pmatrix}\mathbf{x}_t\\ \mathbf{z}_t\end{pmatrix} = A\begin{pmatrix}\mathbf{x}_{t-1}\\ \mathbf{z}_{t-1}\end{pmatrix} + \mathbf{n}_t,\ A\ := \begin{pmatrix}B & C\\ D & E\end{pmatrix}. \tag{10}$$

Assuming a random process model with hidden components, this work further introduces the non-Gaussian independent noise assumption, allowing the probability transition matrix of the statistical model to have causal identifiability. Specifically, it is necessary to assume that the number of hidden time series components does not exceed the total number of observable time series in the system. Under this constraint, by defining an operation known as obtaining the generalized residual R_t, shown in formula (11) where U_1, U_2 denote $d \times d$ arbitrary matrices, one could manage linear combinations towards the noise terms, establishing a connection between R_t and A.

$$R_t(U_1,\ U_2) = \begin{pmatrix}I\\ -U_1\\ -U_2\end{pmatrix}^T \begin{pmatrix}\mathbf{x}_t\\ \mathbf{x}_{t-1}\\ \mathbf{x}_{t-2}\end{pmatrix}. \tag{11}$$

Algorithms then transform the analysis of random processes with hidden components into a parameter estimation problem applying the OICA (overcomplete ICA) and the Variational Estimation (VE), to estimate the parameters of the mixed Gaussian model concerning the noise terms, including B and C, from observable random processes.

Finally, in terms of model unification and generalization, Peters et al. in related work [52] applied the independent noise assumption to SCM in time series, collectively referred to as TiMINo (Time Series Models with Independent Noise). They provided a broader theoretical proof of identifiability for such FCM, along with practical solution based on regression and independence tests. Formula (12) defines the TiMINo model:

$$x_t^i\ := f_i\left(\mathbf{pa}_{t-p}^i, ..., \mathbf{pa}_{t-1}^i, \mathbf{pa}_t^i\right) + n_t^i, \tag{12}$$

where $\mathbf{pa}_t^i \subset \mathbf{x}_t^{N\setminus\{i\}}$, $\mathbf{pa}_{t-p}^i \subseteq \mathbf{x}_{t-p}^N$, $p > 0$. From formula (12), it can be seen that, unlike the VAR-LiNGAM under linear models, TiMINo encompasses the effects of instantaneous and delayed causal functions within a unified generation process, capable of eliminating them in practice through nonlinear regression. For temporal latent confounders, TiMINo interprets the edges whose direction cannot be determined by the algorithm as being influenced by temporal latent confounders.

Compared to the Markov equivalence class (MEC) problem that cannot be resolved by the methods based on CIT introduced in Chapter 2.1.1, a significant advantage of methods based on SCM is that they can uniquely identify causal structures by adding additional assumptions. Specifically, in causal discovery based on incomplete data, methods based on SCM can avoid certain types of temporal latent variable effects by establishing special types of additive models, making them more efficient and accurate in specific scenarios. Meanwhile, under the guarantee of specific causal function identifiability, related methods can practically employ techniques such as Expectation-Maximization (EM) or Variational Estimation (VE) to address latent variable issues. However, a general drawback of methods based on SCM lies in the limitations imposed by function types, making it difficult to cope with more complex nonlinear function types in reality. Additionally, the assumptions regarding data generation forms are often violated when sample sizes are insufficient, leading to biases in function-based regression methods. In the traditional Bayesian network modeling process, score-function-based methods are a classic approach for structural learning from observational data. The next section will provide a detailed discussion of this method and its connections with methods based on CIT and SCM.

2.1.3. Methods Based on Score Functions

In time series analysis, different causal networks correspond to different Dynamic Bayesian Networks (DBN), and methods based on score functions (SF) select the optimal temporal model based on quantifiable evaluation scores within the search space. Therefore, as shown in equation (13), the core idea of methods based on SF for searching the best dynamic Bayesian network $\mathcal{G}^*$ depends on the design and selection of the SF and the search algorithm.

$$\mathcal{G}^* := \arg\min_{\mathcal{G}} S(\mathcal{D},\ \mathcal{G}_{\mathcal{D}}). \tag{13}$$

Here, $\mathcal{D}$ represents the observational dataset, and S denotes the SF under this search strategy. SF are often divided into two main categories: likelihood score functions based on an information-theoretic perspective, and Bayesian score functions based on prior models fusion. The former mainly includes Bayesian Dirichlet equivalent (BDe) scores, K2 scores, etc.; the latter mainly includes Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), etc. Regarding search strategies, due to the need to consider a vast number of graph structure combinations, finding a globally optimal network through combinatorial optimization is known to be an NP-hard problem.

In light of the inherent flaws in the combinatorial optimization approach of traditional score function-based methods, recent work [50] has proposed the DyNotears (Dynamic Notears) model, building on the existing Notears model (Non-combinatorial Optimization via Trace Exponential and Augmented Lagrangian for Structure learning) [76]. Based on the SVAR model, its identifiability is guaranteed by two assumptions: the non-Gaussianity of the noise term, or the noise term following a standard Gaussian distribution with homoscedastic properties [10].

Specifically, from equation (14), the DyNotears model can be interpreted from the perspective of Structural Equation Models (SEM) regarding the parameter matrices in SVAR or DBN models.

$$\begin{cases} \mathbf{x}_t = W_t\, \mathbf{x}_t + \sum_{p=1}^{P} W_{t-p}\mathbf{x}_{t-p} + \mathbf{n}_t. \\ \mathbf{x}_t = W_t\, \mathbf{x}_t + A_p\, \mathbf{y}_p + \mathbf{n}_t. \end{cases} \tag{14}$$

Here, W represents the causal matrix of an acyclic graph, and $A_p = [W_{t-1} \mid ... \mid W_{t-p}]$ represents a matrix of size $d \times pd$ at the current lag p, while $\mathbf{y}_p = [\mathbf{x}_{t-1} \mid ... \mid \mathbf{x}_{t-p}]^T$ represents data of size $pd \times n$ at the current lag p. Therefore, for all time slices, based on the least squares loss $\ell(\cdot)$ and the $\ell - 1$ regularization norm, the constrained optimization problem under the DyNotears model can be initially expressed as

$$\min_{\mathbf{W},\,\mathbf{A}}\ f(\mathbf{W}, \mathbf{A})\ \ s.t.\ \mathbf{W} \in DAGs. \tag{15}$$

Additionally, by introducing the smooth function, $h(\mathbf{W}) = tr(\exp(\mathbf{W} \circ \mathbf{W})) - d$ (as done in the Notears model), it can be understood that $h(\mathbf{W}) = 0$ if and only if $\mathbf{W}$ is acyclic. Thus, the DyNotears model can ultimately be represented as the constrained optimization problem shown in equation (16). By smoothly introducing acyclic constraints, the discrete combinatorial optimization problem is transformed into a continuous optimization problem, allowing for numerical solution methods to be applied for approximate resolution. Furthermore, by utilizing convolutional neural networks and prior knowledge, the extended DyNotears method can be adapted for both linear and nonlinear data.

$$\min_{\mathbf{W},\,\mathbf{A}} \; f(\mathbf{W}, \mathbf{A}) \;\; s.t.\; h(\mathbf{W}) = 0. \tag{16}$$

Overall, methods based on SF, through DyNotears' algebraic perspective, convert a combinatorial optimization problem into the smooth one, successfully extending these methods to high-dimensional time series. This has inspired many subsequent researchers to make further improvements based on continuous optimization problems. However, it should also be noted that there are controversies regarding issues, such as data normalization and the convergence of numerical approximation methods (e.g., augmented Lagrangian)[53].

When assuming the presence of unobserved variables, the classic Structural Expectation-Maximization (SEM) algorithm in traditional dynamic Bayesian network structure learning is the standard algorithm for solving probabilistic networks in incomplete observational data [18]. The Structural Expectation-Maximization algorithm, combined with Bayesian SF based on prior model fusion, can be used for structural learning of dynamic Bayesian networks. Under the assumptions of first-order Markov and stationarity, its core idea is to decompose the dynamic Bayesian network into a prior network and a transition network, followed by iteratively executing optimal parameter estimation for both networks and model selection for different scoring network structures until convergence.

In addition to the classic latent variable structure learning methods for dynamic Bayesian networks mentioned above, SF-based methods are also often combined with TCD methods discussed in Sections 2.1.1 and 2.1.2. For example, the Greedy Equivalence Search (GES) algorithm [12] is a mainstream method for finding optimal graph structures based on the Meek conjecture [46] and dynamic programming. It utilizes the decomposability, score-equivalence, and consistency of score functions to drive heuristic searches through minimal local modifications of the current graph structure. Furthermore, by defining specific constraints, the GES[12] algorithm can be applied to the first stage of the ANLTSM (Additive Non-Linear Time Series Model) [13], where the empirical residuals — containing instantaneous causal effect information obtained through additive regression — will be incorporated into the SF, thus assisting in searching for the correct causal graph structure. Additionally, related work has demonstrated that combining the GES search algorithm with the SVAR-FCI model [43] can also significantly enhance accuracy under finite sample sizes. This hybrid SF-based approach is referred to as the SVAR-GFCI method [43], which combines SF methods during the post-pruning process of SVAR-FCI, heuristically utilizing temporal contextual information and homogeneity assumptions for more efficient edge addition, deletion, and orientation processes.

In summary, by further considering the causal semantics of dynamic Bayesian network structures, we note that the constraint-based methods mentioned in Section 2.1.1, which are based on faithfulness assumptions and conditional independence tests, or the functional class methods based on additive function assumptions mentioned in Section 2.1.1, tend to qualitatively learn directed acyclic graphs, while SF-based methods tend to search all possible generative models from the perspective of probability space, and formulate suitable SF and/or sparsity constraints to quantitatively achieve effective directed acyclic graphs. Hence, researchers have attempted to combine the advantages of both approaches to propose more efficient algorithms; the research approach of SF-based methods is often linked to hybrid TCD algorithm frameworks.

2.1.4. Methods Based on Granger Causality

As another major category widely applied in time series causal inference, Granger causality (G-causality) is one of the classic concepts in TCD [28], often applied in modeling dynamic systems such as economics, neuroscience, and meteorological analysis. It is based on the time series forecasting theory proposed by Wiener, as shown in equation (17), which states that if time series X is determined to be a Granger cause of series X_{t-p}, then the vector autoregressive (VAR) model constructed by both series X and Y will have a smaller prediction error than the VAR model constructed solely from X_{t-p}.

$$\mathbf{x}_t = \sum_{p=1}^{P} \phi(p)\, \mathbf{x}_{t-p} + \mathbf{n}_t. \tag{17}$$

Here, $\phi(p)$ represents the coefficient matrix at lag length p, P is the maximum lag length, and $\mathbf{n}_t$ represents independent noise.

Current work has expanded traditional Granger causality analysis methods from different perspectives, making them applicable in more general scenarios. On one hand, Geweke et al. proposed a conditional Granger causality analysis method (Conditional GC) based on covariance matrix representation and chi-square tests [22][9][6]. In multivariate time series systems, conditional residuals can be calculated for both the full model and the restricted model, and the results can be compared and measured using the conditional Granger causality index (CGCI), thereby extending temporal causal inference to high-dimensional variables. However, high-dimensional time series variables can lead to high computational complexity and false associations among variables. To alleviate this issue, Shojaie et al. introduced a G-causality analysis method based on extended Lasso penalty and variable selection techniques [60].

On the other hand, traditional Granger causality analysis is based on vector autoregressive models, assuming that the causal relationships between variables are linear. However, complex systems in the real world often exhibit intricate nonlinear relationships. In earlier work, Hiemstra and Jones proposed a method based on correlation integral estimation to analyze Granger causality under nonlinearity from a probabilistic measure perspective [32]. Bell et al. approached this from a nonparametric regression perspective, using nonlinear additive models to model time series variables [7]. Furthermore, considering that kernel-based methods can map feature vectors of nonlinear input spaces into high-dimensional linear spaces, Ancona, Marinazzo, and others successively proposed kernel-based RBF-Granger and Kernel-Granger causality analysis methods [44], effectively discovering nonlinear causal relationships. Additionally, as another mainstream nonparametric estimation in time series analysis, transfer entropy (TE) [64][65], based on information-theoretic measures, can measure conditional independence relationships between time series under certain assumptions, allowing for the integration of conditional independence tests and conditional Granger causality analysis frameworks. In recent years, with the widespread application of machine learning methods, especially deep learning technologies, numerous attempts have emerged to explore nonlinear time series relationships based on Granger causality, such as using multilayer perceptron methods [67], minimum prediction information regularization [71], and matrix factorization [73]. Relevant work has been introduced in many reviews.

As mentioned above, numerous model variants based on Granger temporal causality analysis have emerged, with varying research directions and hotspots. This paper, focusing on causal discovery, will briefly describe related GC variants from the perspectives of multivariate time series and nonlinearity, and subsequently introduce the approach of handling latent variables under incomplete data.

Although conditional Granger causality analysis (Conditional G-causality) can handle multivariate time series systems, its causal discovery capability is still affected by spurious associations, including external inputs and hidden confounding variables. To control the interference of hidden confounding variables, Partial Granger causality analysis [29], based on stationary assumptions and autoregressive models (AR), constructs an F-statistic for the covariance matrix of prediction errors inspired by partial correlation statistics. For example, considering the time series pair x^i, x^j under a given set of conditional variables $\mathbf{x}^K$, Partial Granger causality analysis establishes the model as follows:

$$\begin{cases} x_t^j = \sum_{p=1}^{P_1} \phi_{jj}(p)\, x_{t-p}^i + \sum_{p=1}^{P_2} {}_{jk}(p)\, \mathbf{x}_{t-p}^K + \xi_t^j, \\ x_t^i = \sum_{p=1}^{P_3} \phi_{ii}(p)\, x_{t-p}^i + \sum_{p=1}^{P_4} {}_{ik}(p)\, \mathbf{x}_{t-p}^K + \sum_{p=1}^{P_5} \phi_{ij}(p)\, x_{t-p}^j + \xi_t^i. \end{cases} \tag{18}$$

where ξ_t represents the influence of external factors and hidden variables. Based on the above model, covariance matrices are modeled, where S represents the covariance matrix considering only x^i, x^j:

$$S = \begin{pmatrix} var\left(\xi_t^j\right) & cov\left(\xi_t^j, \xi_t^i\right) \\ cov\left(\xi_t^i, \xi_t^j\right) & var\left(\xi_t^i\right) \end{pmatrix} = \begin{pmatrix} s_{jj} & s_{ji} \\ s_{ij} & s_{ii}, \end{pmatrix} \tag{19}$$

and Σ represents the covariance matrix under the given conditional variable set $\mathbf{x}^K$:

$$\Sigma = \begin{pmatrix} var\left(\xi_t^j\right) & cov\left(\xi_t^j, \xi_t^k\right) & cov\left(\xi_t^j, \xi_t^i\right) \\ cov\left(\xi_t^k, \xi_t^j\right) & var\left(\xi_t^k\right) & cov\left(\xi_t^k, \xi_t^i\right) \\ cov\left(\xi_t^i, \xi_t^j\right) & cov\left(\xi_t^i, \xi_t^k\right) & var\left(\xi_t^i\right) \end{pmatrix} = \begin{pmatrix} \Sigma_{jj} & \Sigma_{jk} & \Sigma_{ji} \\ \Sigma_{kj} & \Sigma_{kk} & \Sigma_{ki} \\ \Sigma_{ij} & \Sigma_{ik} & \Sigma_{ii}. \end{pmatrix} \tag{20}$$

Compared to the F-statistic used in conditional Granger causality analysis, $F\left(x_t^j \to x_t^i\right) = \ln\left(\frac{s_{ii}}{\Sigma_{ii}}\right)$, the F-statistic constructed in Partial Granger causality analysis (equation (21)) can better eliminate the influence of external inputs and hidden confounding variables.

$$F\left(x_t^j \to x_t^i\right) = \ln\left(\frac{s_{ii} - s_{ij} s_{jj}^{-1} s_{jj}}{\Sigma_{ii} - \Sigma_{ij} \Sigma_{jj}^{-1} \Sigma_{jj}}\right). \tag{21}$$

On this basis, on one hand, related methods have expanded the application range of Partial Granger causality analysis to nonlinear incomplete observational data through kernel functions. On the other hand, a major drawback of Partial Granger causality analysis is that the proposed F-statistic can sometimes yield negative values, raising concerns about its utility in practical applications. Therefore, recent work [3] has transformed the description of covariance variance relationships in Partial Granger causality analysis into a direct characterization of conditional Gaussian distributions under the assumption of asymmetry in time series models, thus ensuring the monotonic growth of the likelihood function for F-tests, avoiding the inherent negative-value issue of Partial GC. In the frequency domain, other relevant work [15] considers GC analysis using partial directed coherence (PDC) to locally analyze network subsets, thereby discovering hidden components in incomplete observational data.

In recent years, the combination of deep learning technology with Granger causality analysis to identify hidden variables has attracted attention from some scholars. The Temporal Causal Discovery Framework (TCDF) [49] is a method architecture based on deep networks with an attention mechanism to learn complex nonlinear temporal causal relationships. TCDF consists of d identical but independent convolutional neural networks (CNNs), each estimating a specific time series and outputting its attention scores. The core idea of TCDF from the perspective of causality is that, if the current time series has a higher attention score for other time series, it will contain more causal information.

Furthermore, this framework validates and distinguishes causal and correlational (latent confounding) relationships through comparisons of network losses under different time series arrangements.

$$(p_{ji} = p_{ij} = 0) \wedge (p_{kj} = p_{ki}). \tag{22}$$

Under the basic assumption of temporal precedence, as shown in equation (22), TCDF estimates the time steps p (lag causal effects) of time series, allowing it to infer that, time series x_i, x_j with equal delays are influenced by an instantaneous hidden confounding variable x_k. However, the main disadvantage of TCDF is the difficulty in tuning the network's hyperparameters.

The method proposed by combining deep networks with Granger causality analysis [47] uses generative networks based on autoencoders to model and recover temporal hidden confounding factors. This work primarily constructs a generative model between temporal hidden confounding variables and observed variables, sampling from the posterior probability estimates of the hidden confounding variables to correct the original Granger causality analysis. Combining the following equations, this specific method models temporal data using gated recurrent units (GRU) and designs a Temporal Causal Variational Autoencoder (TC-VAE) architecture to estimate the posterior probability q_{θ_z} of hidden confounding variables. However, this method requires further refinement to achieve ideal results.

$$GC\left(x^j \to x^i \mid \mathbf{x}^{V\setminus\{i,j\}},\ \hat{\mathbf{z}}\right),\quad \hat{\mathbf{z}} \sim q_{\theta_z}\left(\hat{\mathbf{z}} \mid x^j, x^i, \mathbf{x}^{P\subset V}\right). \tag{23}$$

Other relevant work [74] assumes that proxy variables selected from observed variables can effectively recover the joint probability distribution, including hidden confounding variables, and effectively encodes the complex generative function using deep network technology, extending nonlinear Granger causality discovery methods to scenarios with hidden confounding variables. Based on the following nonlinear autoregressive model (NAR):

$$\mathbf{x}_t = f(\mathbf{x}_{t-p},\ \mathbf{z}_{t-p}) + \mathbf{n}_t,\ p > 0. \tag{24}$$

The NAR first encodes the hidden confounding factors through proxy variables using multilayer perceptrons (MLPs), constructing the full model: $x_t^i = \hat{f}\left(\mathbf{x}_{t-p}^i,\ x_{t-p}^j,\ \mathbf{z}_{t-p}\right)$, and the corresponding restricted model: $x_t^i = \hat{f}\left(\mathbf{x}_{t-p}^i,\ \mathbf{z}_{t-p}\right)$. Then, it uses nonlinear Granger causality analysis based on recurrent neural networks to conduct dual-decoder testing on the recovered joint probability distribution. As shown in equation (25), if the hidden confounding variables obtained through encoding can reduce the prediction error of the observed time series in Granger causality analysis, one can infer the existence of hidden confounding variables, and thus deduce their causal direction towards observed variables.

$$x^j \xrightarrow{GC} x^i \Leftrightarrow \left[x_t^i - g_1\left(\mathbf{h}_{1:t}^{(x^j)}, g_2\right)\right]^2 < \left[x_t^i - g_2\left(\mathbf{h}_{1:t}^{(x^i)},\ \mathbf{h}_{1:t}^{(\mathbf{z})}\right)\right]. \tag{25}$$

Where $\mathbf{u} \subset \mathbf{x}$, $\mathbf{z} \sim \mathcal{N}\left(\mu\left(\mathbf{h}_{1:t}^{(\mathbf{u})}\right),\ \sigma^2\left(\mathbf{h}_{1:t}^{(\mathbf{u})}\right)\right)$, $\mathbf{h}_t^{(\cdot)}$ represents the hidden state sequence in the GRU unit that contains temporal information, and g represents a nonlinear function approximated by MLPs.

In summary, methods based on Granger causality analysis have a wide range of practical applications and a rich variety of model variants in time series analysis and inference. However, they still heavily rely on the assumption of Causal Sufficiency[61]. In the context of incomplete data, while research tends to design more unbiased statistics to assist Granger causality analysis and avoid erroneous inferences, or to leverage the powerful nonlinear fitting capabilities of deep network technologies to model hidden confounding time series variables, the main issue remains how to better integrate these techniques with more classical and orthodox causal paradigms.

2.2. Causal Discovery Algorithms on Time-Series Applications with Missing Data

In the aforementioned content, this paper primarily discusses theoretical approaches to temporal causal relationships in the presence of unobserved variables. For completeness, this section briefly outlines the issues and challenges in TCD from a practical application perspective, specifically the problems of sample randomness, or non-uniform sampling, typically caused by physical constraints in real-world scenarios. Additionally, considering the prevalence of non-stationary time series data in practical applications, this paper attempts to introduce the relationship between non-stationary effects and the TCD, from the perspective of incomplete data, or external unobserved variables.

2.2.1. Application One: Time-Series Data with Non-Uniform Sampling

For time series with causal relationships, the true causal frequency is unknown; it must be substituted with a fixed sampling frequency, supplemented by subsampling techniques. However, the accompanying decrease in data resolution during the subsampling process undermines the original VAR model assumptions, thereby weakening the practical application of methods such as Granger causality analysis. Research related to subsampling indicates that, under specific assumptions, the subsampled sequence (based on k time steps) $\tilde{\boldsymbol{X}}_t = (\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_{1+k}, ...\tilde{\boldsymbol{x}}_{1+(T-1)k})$ retains structural identifiability compared to the original sequence $\boldsymbol{X}_t = (\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_T)$. Specifically, equation (26) reflects the original VAR model and the subsampled model expanded over k time steps. This work demonstrates that the estimation bias for A^k derived from subsampled data arises from the omission of certain causal relationships of A after aggregation. Consequently, the study proposes finding an appropriate Subsampling Representation Model (A', E', k) to approximate the relationships and theoretically analyzes the connections between A' and A based on assumptions of linear stationarity and non-Gaussian noise.

$$\begin{cases} \boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + E_t, \\ \tilde{\boldsymbol{X}}_t = A^k\tilde{\boldsymbol{X}}_{t-1} + \sum_{l=0}^{k-1} A^l E_{1+tk-l}. \end{cases} \tag{26}$$

In practical applications, this work employs a Gaussian mixture model to model noise aggregation under the assumption of non-Gaussian noise. Given a suitable subsampling factor k, it introduces the Non-Gaussian Expectation-Maximization (NG-EM) algorithm and a more efficient Non-Gaussian Mean-Field (NG-MF) algorithm based on variational inference.

From a broader perspective, methods for data interpolation have shown more stable performance in addressing inherent issues of missing data due to sample randomness or non-uniform sampling, particularly propelled by advancements in graph neural networks (GNNs) and deep learning. Unlike traditional methods that directly use Gaussian processes for time series imputation, the CUTS (Causal discovery from irregUlar Time-Series) model proposes an iterative, non-sequential mutually boosting model. This method iteratively enhances missing data imputation using causal structure information embedded in specialized GNNs, and employs Granger causality analysis as a sparse constraint criterion for further causal graph inference. In multivariate time series analysis, related work has proposed a root cause analysis framework that iteratively combines GNNs and graph structure learning, enabling causal inference of failure relationships among components even in the presence of missing data.

2.2.2. Application Two: Time-Series Data with Non-Stationary Effects

Non-stationary phenomena in time series, where causal model mechanisms change over time, can be interpreted as being influenced by unobserved latent confoundings related to timestamps. Consequently, related work[34] has introduced a more broadly applicable time-dependent causal function model (Time-dependent FCM) as shown in equation (27).

$$\begin{cases} \mathbf{x}_t := \sum_{p=1}^{P} B_p \mathbf{x}_{t-p} + B_0 \mathbf{x}_t + G_t + \mathbf{n}_t, \\ \mathbf{x}_t := f_i\left(\mathbf{x}_{t-p}^N,\ \mathbf{x}_t^N,\ G_t\right) + \mathbf{n}_t. \end{cases} \tag{27}$$

This model[34] considers both linear and nonlinear non-stationary scenarios, and its estimations can be transformed into specific Gaussian process regression problems, thus being referred to as Gaussian Process Models (GP-Model). To capture the uncertainty of incomplete observation systems, the influence of latent variables on observed variables is modeled by a smooth function G_t. Furthermore, considering the practical solution process, the study indicates that the algorithm can assign Gaussian process priors (GP prior) to each time slice, reflecting the causal effect intensity and the influence of latent variables, as shown in equation (28). This allows the Gaussian process model to progressively capture the level of causal influence over time.

$$G_t \sim GP(\mu(t),\ K(t,t)). \tag{28}$$

It is important to note that the establishment and solution of this model fundamentally rely on the assumptions of non-Gaussian independent noise or nonlinear functions. Therefore, in practical estimation, it utilizes the two-stage solving algorithm framework of the VAR-LiNGAM model introduced in section 2.1.2 to estimate the time-delayed and instantaneous causal relationships of the causal function model that includes timestamps as latent confounding variables.

Non-stationary effects are also accompanied by changes in the sampling environment of time series in practical applications. The BackShift model[54], proposed in related work, introduces a causal model that incorporates information about shift interventions in changing environmental domains. It is noteworthy that this model is constrained under linear cyclic conditions, motivated by the consideration that different sampling environments may yield different intervention distributions. Specifically, under assumptions of independence between random interventions and noise terms containing latent variables, the study defines the Cycle Product regarding the changes in causal matrices across different environments. This distinction between cyclic and acyclic graphs enables the transformation of the solution of linear cyclic models into an optimization problem that minimizes the spectral radius of the causal connection matrix.

Through the aforementioned introduction, it is evident that the changes in structural generation processes in non-stationary systems can be viewed as mechanism changes or distribution shifts induced by external latent variables or effects. Related work has demonstrated that the information inherent in non-stationarity aids in causal discovery. The corresponding algorithmic framework CD-NOD (Constraint-based causal Discovery from heterogeneous/NOnstationary Data)[35] can fully utilize the information from mechanism changes or distribution shifts during causal skeleton learning, causal direction determination, and non-stationary representation identification. The main idea is to use surrogate variables to model potential timestamp changes or domain shifts in non-stationary systems, and to integrate causal mechanism invariance theory from a broader perspective — to propose a more general non-parametric causal discovery framework applicable to non-stationary time series with window segmentation independence.

3. Case Studies and Evaluation

This section introduces case studies for temporal causal discovery (TCD), including time-series datasets with incomplete circumstance and common evaluation metrics. By applying TCD algorithms to the corresponding datasets, generalized nonparametric causal discovery estimates the causal graph as the statistical learning objective.

Thus, this section will also briefly introduce the concepts and significance of commonly used assessment metrics in current causal research, focusing on the comparison and trade-off between the estimated causal graph and the true causal graphs.

3.1. Dataset

Datasets used in TCD are generally divided into two categories: simulated data and real world data. **Given that the former can be easily influenced by subjective parameter adjustments, we focuses on introducing real datasets with publicly available causal graphs (ground truth),** briefly summarizing how related work applies these datasets to TCD based on incomplete data.

Real time series datasets with publicly ground truth cover research questions in various fields, including policy, economics, and climate science. The corresponding datasets and experimental processing methods are then listed as follows:

- **Average Daily Discharges of Rivers:** This dataset concerns the time series data of the flow of tributaries of the Danube River, collected by the Bavarian Environmental Agency at different sites established in the watershed. The upstream and downstream relationships determined by natural geographic locations between the sites can serve as the true causal relationships of flow impacts between different regions. Related work [21] considers that the latent confounding relationships in this application context generally have time delays, thus setting different delay parameters (time window sizes) to test the recall rate of identifiably instantaneous causation. Meanwhile, related work [74] introduces data of other representative site flow as proxies for latent confounding variables to validate the effectiveness of TCD methods based on incomplete data. Additionally, it is important to note that this dataset is affected by extreme weather in reality, which often violates the stationary assumption, thereby may impact the application of TCD algorithms [21].
- **Political Economy on Capital Taxation Rates:** The dataset includes twelve variables of interest to economists, such as tax rates, per capita GDP, and national policies, over nearly two decades, which can be used for TCD, especially focusing on the causal impact of other variables on tax policy adjustments. Although this dataset does not contain ground truth, research work [43] qualitatively analyzes the temporal causal graphs obtained by the algorithm within appropriate confidence intervals, pointing out the potential impact of temporal latent confounding factors on the associations between various economic factors, leading to spurious correlations.
- **Stock Markets:** This dataset is based on the Fama-French three-factor model concerning factors influencing stock returns in financial market[17], [39], containing information about financial investment portfolios over a period of 4000 days. Related work [49] removes some confounding variables from its underlying domain knowledge network (causal network) to evaluate the effectiveness of related methods in TCD based on incomplete data.
- **Temperature Ozone:** Related work [27] explores the binary causal effect identification relationship between ozone and temperature in this time series dataset. The research conducts subsampling to reduce data resolution, in an effort to compare the recovery degree of causal matrix strength by TCD methods on subsampled data under different sampling factors.

- **Commodities Price Time Series:** The time series vector includes price fluctuation data of three products: cheese, butter, and milk, from January 1968 to April 2014. Linear causal matrix estimation is performed on both the total time series data composed of the three products, and the time series data with one product omitted (composed of the other two products). The analysis process of this dataset is applied in related work [19], but it is necessary to conduct additional model checks on the data to ensure that the assumptions of relevant temporal latent variable causal discovery algorithms are satisfied (e.g., Gaussianity of noise).
- **Temperature in House:** This time series data records the temperature in different areas near a house in the suburbs of Germany every hour. Common sense suggests that the external temperature of the house does not affect the internal regions. Therefore, based on prior knowledge, the temperature recorded in an external area of the house can be viewed as a latent confounding variable, estimating the causal structure between areas and their changes over time under conditions such as whether someone is living there (causing temperature changes due to electric heating devices). Specific analyses and applications can be found in related work [52], [34], [27].
- **Financial Time Series:** Related work [34], [54] analyzes time series containing different stock indices, considering that temporal data in the financial field exhibits characteristics that change with the environment. Although different environments act as latent variables, they contain intervention information. The research can estimate intervention intensity and has made good predictions on this dataset based on intervention differences.

3.2. Assessment Metrics

Before introducing typical metrics, **it is notable to outlines basic evaluation modules by analogizing classification metrics (in realms of machine learning) to causal structures discovery**: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Common evaluation metrics defined around the comparison and trade-off between the estimated causal graph and the true causal graph (ground truth) will then be listed in the following.

Specifically, TP and FP can be seen as a measure of the causal relationships identified in the estimated causal graph. TP can be analogized to the total number of variable pairs (causal edges) identified in the estimated causal graph that are consistent with the causal relationships present in the true causal graph, quantifying the correctly estimated causal relationships between variables. FP can be analogized to the total number of incorrectly identified variable pairs in the estimated causal graph that do not have causal relationships in the true causal graph. TN and FN can be seen as measures of the causal relationships not identified in the estimated causal graph. TN can be analogized to the total number of variable pairs in the estimated causal graph that are not identified but are consistent with the absence of causal relationships in the true causal graph, quantifying the independence relationships correctly estimated between variables. FN can be analogized to the total number of variable pairs (causal edges) that are missed in the estimated causal graph but have causal relationships in the true causal graph.

- **Precision** refers to the proportion of correctly identified causal relationships in the estimated causal graph among all estimated causal relationships. In other words, the higher the precision in the estimated causal graph, the more reliable the identification of causal relationships between variable pairs in the estimated causal graph.

$$Precision = \frac{TP}{TP + FP}. \tag{29}$$

- **Recall** refers to the proportion of correctly identified causal relationships in the estimated causal graph among all true causal relationships. Alternatively, the higher the recall in the estimated causal graph, the more it can encompass the causal relationships over the true structure.

$$Recall = \frac{TP}{TP + FN}. \tag{30}$$

- **F_1 Score** is the harmonic mean of precision and recall, integrating the advantages of both and serving as an overall measure of the effectiveness of causal discovery.

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{31}$$

- **True Positive Rate** refers to the proportion of correctly identified causal relationships in the estimated causal graph among all true causal relationships (aka the recall).

$$TPR = \frac{TP}{TP + FN}. \tag{32}$$

- **False Positive Rate**, also known as specificity, refers to the proportion of incorrectly identified causal relationships in the estimated causal graph among all true non-existent causal relationships. A higher false positive rate indicates that the estimated causal graph deviates more from the true causal graph, and tends to classify causal relationships that do not exist in reality as existent.

$$FPR = \frac{FP}{FP + TN}. \tag{33}$$

- **ROC Curve** (Receiver Operating Characteristic Curve) reflects the relationship between the false positive rate on the horizontal axis and the true positive rate on the vertical axis under a given classification threshold $\boldsymbol{k}$, denoted as (FPR_k, TPR_k). When the true positive rate equals the false positive rate ($FPR = 0.5, TPR = 0.5$), it forms a baseline l on the coordinate axis representing random guessing. The degree to which the ROC curve deviates from l is characterized by the Area Under the ROC Curve (AUROC); a larger AUROC value indicates a greater degree of deviation from the random guessing baseline l, reflecting that the causal discovery algorithm has good predictive performance compared to random guessing (under suitable thresholds).

$$AUROC = \int_{\boldsymbol{k}} TPR_k \, d(FPR_k). \tag{34}$$

- **Structural Hamming Distance** measures the gap between the estimated causal graph and the true causal graph from the perspective of the accuracy of estimating causal edges. This metric quantifies the differences by summing the missing edges (ME), extra edges (EE), and reverse edges (RE) in the estimated graph compared to the true causal graph.

$$SHD = ME + EE + RE. \tag{35}$$

4. Discussion

Under the ideal conditions of basic temporal precedence assumptions and temporal stationarity assumptions, **we in this review categorizes various temporal causal discovery (TCD) methods for handling incomplete data into the three of main paradigms**: algorithms based on *causal Markovity and causal faithfulness assumptions*, algorithms based on *independent noise assumptions*, and algorithms based on *Granger predictive theory*. Table 2 (in the next page) presents typical TCD methods based on incomplete data corresponding to these different assumptions.

Given the assumptions of causal Markovity and causal faithfulness, it means that the graphical structure is compatible with probability decomposition. The paradigm for handling temporal latent variables is represented by methods based on conditional independence tests (CIT). Based on temporal precedence, where causal arrows always move in the direction of time, and the conditional independence relationships among variables reflected by maximal ancestral graphs (MAGs), pairwise variables that cannot entail independence by statistical constraints will imply either the existence of causal relationships, or the influence of latent confounding variables. In the context of time series, since not

Table 2
Paradigms for processing TCD (Temporal Causal Discovery) with incomplete data

Primary Assumptions as to Causal Identifiability			Algorithms(Models/Ideas)
Temporal Precedence Assumption Temporal Stationarity Assumption	Markov Assumption Faithfulness Assumption		[21],[16],[45],[13],[2],[35]
	Independent Noise Assumption	Non-Gaussianity Non-linearity	[19], [34], [35], [27] [13], [52], [34],[35]
	Granger Predictive Theory	Surrogate/Latent Space Adequacy	[29],[3],[49] [47], [74]

considering the autocorrelation or long-range causal effects between variables can introduce latent confounding effects, under the assurance of temporal stationarity, a heuristic search through extended time series to control the selection of condition sets for conditional independence testing can help avoid such latent confounding effects.

When the independent noise assumption holds, independence tests and statistical analysis can be applied to observational data, with the paradigm for handling temporal latent variables represented by methods based on structural causal models (SCM). When the noise follows non-Gaussian distributions and the functions causal model (FCM) remain linear, non-Gaussian independent noise can be utilized to optimize the temporal transfer matrix that introduces latent components. If the causal coefficients corresponding to the latent components in the matrix are non-zero, the influence and structure of the latent components can be inferred. When the FCM may possess a nonlinear and additive noise form, nonlinearity could yield causal identifiability effects similar to those of non-Gaussian noise. The fact that, latent confounding can disrupt the independent noise assumption, can be tested by examining whether variables still exhibit dependence after eliminating the influences of temporal autoregression and other variables, thereby judging the existence of latent confounding variables.

Under the support of Granger predictive theory, the paradigm for handling temporal latent variables is reflected in different variants of Granger causal analysis methods. In the presence of latent variable influence, traditional methods aim to avoid potentially erroneous inferences or mitigate incorrect inferences by statistically correcting the covariance based on conditional Granger predictive theory; this correction can construct more robust Granger statistics. In recent years, deep learning techniques under unsupervised learning have provided possibilities for fitting complex functional relationships. Under the assumptions of latent space adequacy or the sufficiency of latent (confounding) variable spaces, temporal latent confounding variables can typically be recovered through an autoencoder (AE) framework, either via surrogate variables or directly through global observational variables. According to the concepts of Granger predictive theory, whether the recovered latent confounding variables can reduce the model's prediction error can be used to judge the existence of latent confounding structures.

Ultimately, it should be noted that another fundamental theoretical method for TCD, based on scoring functions, often combines and coordinates the aforementioned assumptions within the same framework. Moreover, considering that TCD based on incomplete data in practical applications often struggles to meet stationarity assumptions, causes of such issues can be attributed to latent confounding variables over time slices, leading researchers to assign Gaussian priors about the latent confounding variables to each time slice. This transforms the problem into one of solving specific Gaussian Process Regression and independence tests, fundamentally relying on the non-Gaussian noise assumption or nonlinearity mentioned above. Therefore, the paradigms for handling TCD based on incomplete data in these scenarios are fundamentally consistent with the aforementioned methods.

5. Conclusion

Based on the current state of the field, two aspects that can be observed and inferred are: **(1)** Existing methods for temporal causal discovery (TCD) based on incomplete data tend to explicitly parallel the promotion of conventional non-temporal methods. However, this promotion often aims to achieve correspondingly algorithmic variants through the strategies of local combination and optimization. That is, preemptively resolving the time dimension in TCD problems (i.e., the influence of delay effects), in an effort to transform the problem into the superposition of multiple non-temporal issues (i.e., partitioning datasets or implementing multi-step causal discovery strategies). Therefore, how to genuinely utilize time dimension information based on the characteristics of time series problems, to solve the challenges of TCD based on incomplete data from a more holistic perspective, is one of the main challenges facing future research. **(2)** It is undoubted that widely used TCD methods based on Granger causal analysis and prediction have rich historical reasons and model variants, especially with the sweeping influence of deep learning technologies. Temporal variables or predictive functions in Granger causal analysis, including latent confounding temporal factors, can often be constructed and fitted by complex and powerful deep networks. However, the controversial argument lies in their interpretation regarding the actual causal implications (e.g., intervention or structural causal model representation). Therefore, methods based on Granger causality still need to be explored in future research based on specific practical application scenarios.

In summary, although the existence of the temporal dimension and different types of temporal latent confounding variables complicate and challenge the task of causal discovery, this review hopes to provide a reference perspective on how TCD methods can more reasonably and cleverly leverage (real-world) temporal information to assist causal inference, making the results of TCD based on incomplete data more convincing and reliable.

Acknowledgments

During the research internship, Xuanzhi Chen appreciate the opportunity given by the DMIR lab for managing a research review writing (2023 spring - 2023 fall). The topic, temporal causal discovery with incomplete data, serves as an extension of Wei Chen's previous research work (*A Survey on Non-Temporal Series Observational Data based Causal Discovery, 2017*). With lab meeting discussion, Ruichu Cai helped defined the core perspective of this review at Xuanzhi's work presentation.

References

[1] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering–a decade review". In: *Information systems* 53 (2015), pp. 16–38.

[2] Bryan Andrews, Peter Spirtes, and Gregory F Cooper. "On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 4002–4011.

[3] Takashi Arai. "A new Granger causality measure for eliminating the confounding influence of latent common inputs". In: *arXiv preprint arXiv:1908.03867* (2019).

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

[5] Lionel Barnett, Adam B Barrett, and Anil K Seth. "Granger causality and transfer entropy are equivalent for Gaussian variables". In: *Physical review letters* 103.23 (2009), p. 238701.

[6] Adam B Barrett, Lionel Barnett, and Anil K Seth. "Multivariate Granger causality and generalized variance". In: *Physical Review E* 81.4 (2010), p. 041907.

[7] David Bell, Jim Kay, and Jim Malley. “A non-parametric approach to non-linear causality testing”. In: *Economics Letters* 51.1 (1996), pp. 7–18.

[8] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. “Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 884–895.

[9] Andrea Brovelli et al. “Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality”. In: *Proceedings of the National Academy of Sciences* 101.26 (2004), pp. 9849–9854.

[10] Peter Bühlmann, Jonas Peters, and Jan Ernest. “CAM: Causal additive models, high-dimensional order search and penalized regression”. In: (2014).

[11] Yuxiao Cheng et al. “CUTS: Neural Causal Discovery from Irregular Time-Series Data”. In: *arXiv preprint arXiv:2302.07458* (2023).

[12] David Maxwell Chickering. “Optimal structure identification with greedy search”. In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[13] Tianjiao Chu, Clark Glymour, and Greg Ridgeway. “Search for Additive Nonlinear Time Series Causal Models.” In: *Journal of Machine Learning Research* 9.5 (2008).

[14] Diego Colombo et al. “Learning high-dimensional directed acyclic graphs with latent and selection variables”. In: *The Annals of Statistics* (2012), pp. 294–321.

[15] Heba Elsegai et al. “Granger Causality Analysis in the Presence of Latent Confounders”. In: (2014).

[16] Doris Entner and Patrik O Hoyer. “On causal discovery from time series data using FCI”. In: *Probabilistic graphical models* (2010), pp. 121–128.

[17] Eugene F Fama and Kenneth R French. “The cross-section of expected stock returns”. In: *the Journal of Finance* 47.2 (1992), pp. 427–465.

[18] Nir Friedman et al. “Learning belief networks in the presence of missing values and hidden variables”. In: *Icml*. Vol. 97. July. Berkeley, CA. 1997, pp. 125–133.

[19] Philipp Geiger et al. “Causal inference by identification of vector autoregressive processes with hidden components”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1917–1925.

[20] Andreas Gerhardus. “Characterization of causal ancestral graphs for time series with latent confounders”. In: *arXiv preprint arXiv:2112.08417* (2021).

[21] Andreas Gerhardus and Jakob Runge. “High-recall causal discovery for autocorrelated time series with latent confounders”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12615–12625.

[22] John Geweke. “Measurement of linear dependence and feedback between multiple time series”. In: *Journal of the American statistical association* 77.378 (1982), pp. 304–313.

[23] Will Glad and Tom Woolf. “Path Signature Area-Based Causal Discovery in Coupled Time Series”. In: *Causal Analysis Workshop Series*. PMLR. 2021, pp. 21–38.

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

[25] Chang Gong et al. “Causal Discovery from Temporal Data: An Overview and New Perspectives”. In: *arXiv preprint arXiv:2303.10112* (2023).

[26] Mingming Gong et al. “Causal discovery from temporally aggregated time series”. In: *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*. Vol. 2017. NIH Public Access. 2017.

[27] Mingming Gong et al. “Discovering temporal causal relations from subsampled data”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1898–1906.

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

[29] Shuixia Guo et al. “Partial Granger causality—eliminating exogenous inputs and latent variables”. In: *Journal of neuroscience methods* 172.1 (2008), pp. 79–93.

[30] Eduardo Hariton and Joseph J Locascio. “Randomised controlled trials—the gold standard for effectiveness research”. In: *BJOG: an international journal of obstetrics and gynaecology* 125.13 (2018), p. 1716.

[31] Uzma Hasan, Emam Hossain, and Md Osman Gani. “A Survey on Causal Discovery Methods for Temporal and Non-Temporal Data”. In: *arXiv preprint arXiv:2303.15027* (2023).

[32] Craig Hiemstra and Jonathan D Jones. “Testing for linear and nonlinear Granger causality in the stock price-volume relation”. In: *The Journal of Finance* 49.5 (1994), pp. 1639–1664.

[33] Patrik Hoyer et al. “Nonlinear causal discovery with additive noise models”. In: *Advances in neural information processing systems* 21 (2008).

[34] Biwei Huang, Kun Zhang, and Bernhard Schölkopf. “Identification of time-dependent causal model: A gaussian process treatment”. In: *Twenty-Fourth international joint conference on artificial intelligence*. 2015.

[35] Biwei Huang et al. “Causal discovery from heterogeneous/nonstationary data”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 3482–3534.

[36] Aapo Hyvärinen, Shohei Shimizu, and Patrik O Hoyer. “Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 424–431.

[37] Hassan Ismail Fawaz et al. “Deep learning for time series classification: a review”. In: *Data mining and knowledge discovery* 33.4 (2019), pp. 917–963.

[38] Samantha Kleinberg. “A logic for causal inference in time series with discrete and continuous variables”. In: *Twenty-Second International Joint Conference on Artificial Intelligence*. 2011.

[39] Samantha Kleinberg. *Causality, probability, and time*. Cambridge University Press, 2013.

[40] Lei Li and B Aditya Prakash. “Time series clustering: Complex is simpler!” In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 185–192.

[41] Jason Lines and Anthony Bagnall. “Ensembles of elastic distance measures for time series classification”. In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM. 2014, pp. 524–532.

[42] Lars Lorch et al. “Dibs: Differentiable bayesian structure learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 24111–24123.

[43] Daniel Malinsky and Peter Spirtes. “Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding”. In: 92 (2018), pp. 23–47.

[44] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. “Kernel-Granger causality and the analysis of dynamical networks”. In: *Physical review E* 77.5 (2008), p. 056215.

[45] Atalanti A Mastakouri, Bernhard Schölkopf, and Dominik Janzing. “Necessary and sufficient conditions for causal feature selection in time series with latent common causes”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 7502–7511.

[46] Christopher Meek. “Graphical Models: Selecting causal and statistical models”. PhD thesis. PhD thesis, Carnegie Mellon University, 1997.

[47] Yuan Meng. “Estimating Granger causality with unobserved confounders via deep latent-variable recurrent neural network”. In: *arXiv preprint arXiv:1909.03704* (2019).

[48] Raha Moraffah et al. “Causal inference for time series analysis: Problems, methods and evaluation”. In: *Knowledge and Information Systems* 63 (2021), pp. 3041–3085.

[49] Meike Nauta, Doina Bucur, and Christin Seifert. “Causal discovery with attention-based convolutional neural networks”. In: *Machine Learning and Knowledge Extraction* 1.1 (2019), pp. 312–340.

[50] Roxana Pamfil et al. “Dynotears: Structure learning from time-series data”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 1595–1605.

[51] Judea Pearl et al. “Models, reasoning and inference”. In: *Cambridge, UK: CambridgeUniversityPress* 19.2 (2000).

[52] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. “Causal inference on time series using restricted structural equation models”. In: *Advances in neural information processing systems* 26 (2013).

[53] Alexander Reisach, Christof Seiler, and Sebastian Weichwald. “Beware of the simulated dag! causal discovery benchmarks may be easy to game”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27772–27784.

[54] Dominik Rothenhäusler et al. “BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions”. In: *Advances in Neural Information Processing Systems* 28 (2015).

[55] Jakob Runge. “Causal network reconstruction from time series: From theoretical assumptions to practical estimation”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.7 (2018), p. 075310.

[56] Jakob Runge et al. “Detecting and quantifying causal associations in large nonlinear time series datasets”. In: *Science advances* 5.11 (2019), eaau4996.

[57] Ulrich Schaechtle, Kostas Stathis, and Stefano Bromuri. “Multi-dimensional causal discovery”. In: *twenty-third international joint conference on artificial intelligence*. 2013.

[58] Shohei Shimizu et al. “A linear non-Gaussian acyclic model for causal discovery.” In: *Journal of Machine Learning Research* 7.10 (2006).

[59] Shohei Shimizu et al. “DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model”. In: *Journal of Machine Learning Research-JMLR* 12.Apr (2011), pp. 1225–1248.

[60] Ali Shojaie and George Michailidis. “Discovering graphical Granger causality using the truncating lasso penalty”. In: *Bioinformatics* 26.18 (2010), pp. i517–i523.

[61] Peter Spirtes et al. *Causation, prediction, and search*. MIT press, 2000.

[62] Adolf Stips et al. “On the causal structure between CO2 and global temperature”. In: *Scientific reports* 6.1 (2016), p. 21691.

[63] George Sugihara et al. “Detecting causality in complex ecosystems”. In: *science* 338.6106 (2012), pp. 496–500.

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

[65] Jie Sun, Dane Taylor, and Erik M Bollt. “Causal network inference by optimal causation entropy”. In: *SIAM Journal on Applied Dynamical Systems* 14.1 (2015), pp. 73–106.

[66] Floris Takens. “Detecting strange attractors in turbulence”. In: *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*. Springer. 2006, pp. 366–381.

[67] Alex Tank et al. “Neural granger causality for nonlinear time series”. In: *stat* 1050 (2018), p. 16.

[68] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. “D’ya like dags? a survey on structure learning and causal discovery”. In: *ACM Computing Surveys* 55.4 (2022), pp. 1–36.

[69] Yuyang Wang et al. “Deep factors for forecasting”. In: *International conference on machine learning*. PMLR. 2019, pp. 6607–6617.

[70] Andreas S Weigend. *Time series prediction: forecasting the future and understanding the past*. Routledge, 2018.

[71] Tailin Wu et al. “Discovering nonlinear relations with minimum predictive information regularization”. In: *arXiv preprint arXiv:2001.01885* (2020).

[72] Tianhao Wu et al. “Nonlinear Causal Discovery in Time Series”. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022, pp. 4575–4579.

[73] Chenxiao Xu, Hao Huang, and Shinjae Yoo. “Scalable causal graph learning through a deep neural network”. In: *Proceedings of the 28th ACM international conference on information and knowledge management*. 2019, pp. 1853–1862.

[74] Zexuan Yin and Paolo Barucca. “Deep recurrent modelling of Granger causality with latent confounding”. In: *Expert Systems with Applications* 207 (2022), p. 118036.

[75] Min Zheng and Samantha Kleinberg. “Using domain knowledge to overcome latent variables in causal inference from time series”. In: *Machine Learning for Healthcare Conference*. PMLR. 2019, pp. 474–489.

[76] Xun Zheng et al. “Dags with no tears: Continuous optimization for structure learning”. In: *Advances in neural information processing systems* 31 (2018).