Shylock: Causal Discovery in Multivariate Time Series based on Hybrid Constraints

Shuo Li, Keqin Xu

Institute of Software,Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing 100049, China
{lishuo,xukeqin}19@otcaix.iscas.ac.cn

Jie Liu*, Dan Ye

Institute of Software,Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing 100049, China
University of Chinese Academy of Sciences
Nanjing, Nanjing 211135, China
{ljie,yedan}@otcaix.iscas.ac.cn

Abstract—Causal relationship discovery has been drawing increasing attention due to its prevalent application. Existing methods rely on human experience, statistical methods, or graphical criteria methods which are error-prone, stuck at the idealized assumption, and rely on a huge amount of data. And there is also a serious data gap in accessing Multivariate time series(MTS) in many areas, adding difficulty in finding their causal relationship. Existing methods are easy to be over-fitting on them.

To fill the gap we mentioned above, in this paper, we propose Shylock, a novel method that can work well in both few-shot and normal MTS to find the causal relationship. Shylock can reduce the number of parameters exponentially by using group dilated convolution and a sharing kernel, but still learn a better representation of variables with time delay. By combing the global constraint and the local constraint, Shylock achieves information sharing among networks to help improve the accuracy. To evaluate the performance of Shylock, we also design a data generation method to generate MTS with time delay. We evaluate it on commonly used benchmarks and generated datasets. Extensive experiments show that Shylock outperforms two existing state-of-art methods on both few-shot and normal MTS. We also developed Tcausal, a library for easy use and deployed it on the earthDataMiner platform [1].

Index Terms—Multivariate time series, Causality, Causal discovery

I. Introduction

Time series data can help uncover relationships between variables. Multivariate time series (MTS) data is generated when recording time series from a wide range of sensors.

Existing researches utilize a huge amount of MTS for forecasting. It has seen tremendous applications in the domains of economics, finance, bioinformatics, and traffic [22] [3].

But in recent years, some researchers are more concerned with the causal relationships among the variables in MTS data. By identifying causality, researchers and practitioners can gain a deeper understanding of how changes in one variable affect other variables, and can make more informed decisions and predictions. For example, in recent years, the rapidly treated Arctic sea ice has attracted much attention which is also an important point in the global Sustainable Development Goals

which lay out a comprehensive and ambitious agenda for global development[2]. Knowing the causal relationship between retreated Arctic sea ice with other factors can further help protect the environment. So some researchers struggled to collect these related data, such as the global land degradation rate and the world's groundwater usage rate, to find their causal relationships. These data are extremely difficult to collect [3]. We refer to these time series with a tiny amount of data as few-shot multivariate time series. Furthermore, there usually exists a time delay among variables in MTS data, which further increases the difficulty of finding causal relationships.

Traditional methods rely on human experience to find causal relations, which is time-consuming and error-prone. Some statistical methods [10] [9] [1] [13] or graphical criteria methods rely on a non-trivial combination of probability axioms [12] [25]. However, these methods can not work well on MTS data as they can not characterize time delay features of data or are stuck at idealized assumptions (e.g., the data is without any noise). Causal discovery aims to discover direct cause-effect relationships for both instantaneous and delayed causes [11].

In recent years, benefiting from the development of deep learning techniques, some of the advanced deep learning-based methods, such as TCDF [14], have been proposed to find causal relationships. It can model the time delay by the neural networks to enhance the ability to learn causal relationships. However, these methods depend on a huge amount of data and parameters. In many areas, it may be difficult and infeasible to collect a large amount of data, and the data scale is so small even less than hundreds of items. There are serious data gaps in assessing the aforementioned data. Furthermore, these methods suffer from being over-parametrization and difficult to converge, and difficult to learn the generalized temporal feature expression when applied on few-shot MTS. At the same time, few-shot MTS is always characterized by high noise and time delay, increasing the difficulty of finding the causal relationship. For example, the total number of parameters in TCDF is not less than $O(nnf)$, where n

---

[2] https://sdgs.un.org/
[3] https://blogs.worldbank.org/opendata/are-we-there-yet-many-countries-dont-report-progress-all-sdgs-according-world-banks-new

represents the num of time series (also called variables) and f represents the length of receptive field(almost the same as the length of timesteps). When applied TCDF on few-shot MTS, our experiment proves that it is over-fitting. Detailed information is shown in Section 5.

To fill the gap we mentioned above, we propose **Shylock**, a novel method that can effectively find causal relationships on multivariate time series even on few-shot multivariate time series. Shylock independently models the causal relationship between variables using a neural network. To solve the time delay of variables and reduce the number of parameters exponentially, Shylock uses group dilated convolution in each network to learn a better representation of variables, and a sharing kernel to learn local causal relationships among variables. Then Shylock uses a global loss to obtain the global causal relationships, which are represented by the attention matrix. To identify cyclic causal relationships. Shylock conducts constraints on the global loss and attention and eliminates cyclic causal relationships by DAG.

To evaluate Shylock, we design a lightweight method to generate MTS data with time delay. Based on that, we generate few shot datasets and use Shylock to find causal relationships among the variables. Besides, we also choose a benchmark, FMRI [7], to evaluate the performance of Shylock. The experiment results show that Shylock is effective and efficient in finding causal relationships in MTS data, and significantly outperforms two existing state-of-art methods, NOTEARS and TCDF. Overall, our contributions are as follows:

- To the best of our knowledge, we are the first to emphasize the causal relationship discovery on few-shot multivariate time series. We propose **Shylock**, a neural-network-based method that incorporates hybrid constraints to mine causal relationships on Multivariate Time Series even in limited data scenarios.
- SShylock reduces the number of parameters exponentially and uses hybrid constraints to facilitate information sharing during training and prediction without compromising performance. It addresses time delays in MTS data and minimizes parameterization by employing group dilated convolutions and a shared kernel to learn local causal relationships. Shylock further applies global constraints via a DAG to eliminate cyclic causal relationships, combining local and global constraints to infer causal connections.
- To evaluate the effectiveness of Shylock, we developed a lightweight method to generate MTS data with time delays. Experiments on the generated datasets and a common benchmark show that Shylock outperforms two state-of-the-art methods in identifying causal relationships in MTS data with time delays.

II. INTRODUCTION

Time series data can help uncover relationships between variables. Multivariate time series (MTS) data is generated when recording time series from a wide range of sensors.

Existing researches utilize a huge amount of MTS for forecasting. It has seen tremendous applications in the domains of economics, finance, bioinformatics, and traffic [22] [3].

But in recent years, some researchers are more concerned with the causal relationships among the variables in MTS data. By identifying causality, researchers and practitioners can gain a deeper understanding of how changes in one variable affect other variables, and can make more informed decisions and predictions. For example, in recent years, the rapidly treated Arctic sea ice has attracted much attention which is also an important point in the global Sustainable Development Goals which lay out a comprehensive and ambitious agenda for global development[4]. Knowing the causal relationship between retreated Arctic sea ice with other factors can further help protect the environment. So some researchers struggled to collect these related data, such as the global land degradation rate and the world's groundwater usage rate, to find their causal relationships. These data are extremely difficult to collect [5]. We refer to these time series with a tiny amount of data as few-shot multivariate time series. Furthermore, there usually exists a time delay among variables in MTS data, which further increases the difficulty of finding causal relationships.

Traditional methods for causal discovery depend heavily on human expertise, which is time-consuming and error-prone. Statistical techniques [1], [9], [10], [13] and graphical methods based on probability axioms [12], [25] are often employed for multivariate data. However, they struggle with MTS data due to limitations in capturing time delays or reliance on unrealistic assumptions, such as noise-free data [11].

Benefiting from the development of deep learning techniques, some of the advanced deep learning-based methods, such as TCDF [14], have shown promise in modeling time delays and causal discovery. However, these methods require large amounts of data and parameters, which can be difficult to obtain, especially in fields with small datasets (often fewer than hundreds of samples). Furthermore, these techniques suffer from over-parameterization, poor convergence, and challenges in learning generalized temporal features, particularly when applied to few-shot MTS data, which is often noisy and involves time delays. For instance, TCDF has at least $O(nnf)$ parameters, where n is the number of time series and f is the receptive field length (approximately equal to the number of timesteps). Our experiments show that TCDF tends to overfit when applied to few-shot MTS, as detailed in Section 5.

To fill the gap we mentioned above, we propose **Shylock**, a novel method designed to effectively discover causal relationships in multivariate time series (MTS), even in few-shot settings. Shylock independently models causal relationships between variables using neural networks. To handle time delays and reduce parameter counts exponentially, Shylock employs group dilated convolutions within each network to learn improved representations of variables, while a shared kernel captures local causal relationships. It then uses a

---

[4]https://sdgs.un.org/

[5]https://blogs.worldbank.org/opendata/are-we-there-yet-many-countries-dont-report-progress-all-sdgs-according-world-banks-new

global loss to infer global causal relationships, represented by the attention matrix. To address cyclic causal relationships, Shylock introduces constraints on the global loss and attention, eliminating cycles through a directed acyclic graph (DAG).

To evaluate Shylock, we design a To evaluate Shylock, we developed a lightweight method to generate MTS data with time delays, creating few-shot datasets for causal discovery. We also benchmarked Shylock against FMRI data [7] to assess its performance. Experimental results demonstrate that Shylock is both effective and efficient in identifying causal relationships in MTS data, significantly outperforming two state-of-the-art methods, NOTEARS and TCDF. Overall, our contributions are as follows:

- To the best of our knowledge, we are the first to emphasize the causal relationship discovery on few-shot multivariate time series. We propose **Shylock**, a neural-network-based method that incorporates hybrid constraints to mine causal relationships on Multivariate Time Series even in the case of few-shot series.
- Shylock can reduces the number of parameters exponentially and leverages hybrid constraints to facilitate information sharing during training and prediction without degrading the model performance . To solve time delay in MTS data and release parametrization, Shylock employs group dilated convolutions and a sharing kernel to learn local causal relationships among variables. Shylock conducts global constraints to identify and eliminate cyclic causal relationships using DAG. By combining the local and global constraints, Shylock imposes a global loss to find causal relationships.
- To evaluate the effectiveness of Shylock, we also proposed a lightweight method to generate MTS data with time delay. Based on the generated dataset and commonly used benchmark, experiment results show that, compared with two existing state-of-art methods, Shylock is more effective and efficient in finding causal relationships among MTS data with time delay.

III. PROBLEM STATEMENT

Multivariate Time Series data: Temporal causal discovery between multivariate time series from observational can be formulated as follows:

As shown in Figure 1(a), the dataset $X = \{x_1, ..., x_n\}$ are consisted of n observed time series of the same length t. These time series are also called variables. For sake of brevity, in the following paper, we denoted the times series as variables in the following, and data collected by each observation variable can be aligned in time.
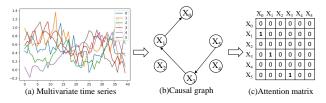

Fig. 1: Temporal causal graph representation.

Causal graph: Then these causal relationships are commonly expressed as a causal graph $\mathcal{G}$, which can be represented as a directed acyclic graph (DAG) according to the common assumption. As shown in Figure 1(b), the vertices are time series $x_k \in X$ and the arrows are direct causal relationships.

Attention matrix: In order to formalize the graphical constraints on DAG and for the convenience of calculations, an attention matrix A is introduced, shown in Figure 1(c). $a_{i,j} >= threshold$ represents the i_{th} time series x_i as the effect and the j_{th} time series x_j as the cause and conversely.

IV. SHYLOCK

This section presents a neural network-based approach for mining causal relationships in multivariate time series (MTS). Figure 2 illustrates the architecture of Shylock, comprising two key steps:

(1) **Attention-based local causal discovery:** Shylock models potential causes for each variable by constructing n sub-convolutional neural networks (CNNs), each targeting a single variable. To address overfitting on few-shot MTS, shared convolution kernels minimize parameter size, while a grouped dilated convolution module captures time-delayed causal effects.

(2) **Adjacent matrix constraint-based global causal discovery:** By combining global constraints with local fitting objectives, Shylock ensures acyclic causal relationships. Attention vectors from sub-CNNs are combined to form an attention matrix, but due to the lack of direct information sharing, the matrix alone does not guarantee acyclicity. Therefore, adjacency matrix constraints are applied to enforce this property.

In the following sections, we describe each step in more detail. For easy description, we formalize some related definitions and describe each step in detail in the following sections.

A. Attention-based local causal discovery with Parameter Sharing

Shylock identifies causal relationships by building CNNs for each variable x_i, as shown in the blue dotted box of Figure 2. The CNNs use two convolution kernel types: ①Grouped Dilated Convolution, which Models time-series features. ②One-dimensional Convolution, which aggregates these features to capture potential causes for the target variable.

Attention matrix: As shown on Figure 2 The matrix A represents attention relationships between variables, defined as $A = \{a_1, ..., a_n\}$, where $a_k = \{a_{1,k}, ..., a_{n,k}\}$ is an $N * 1$ vector representing weights for the k_{th} network. Initially, self-causation weights a_{kk} are preset to α (commonly 0), with others set to 1. During training, the attention matrix adjusts dynamically, ultimately determining the causal relationships based on thresholded attention weights $a_{i,j}$. Further details on the global constraints applied to these networks are discussed in Section IV-B.

Group dilated convolution: Each network aims to discover causal relationships between a target variable and others. Sparse MTS often exhibit long time delays, necessitating a
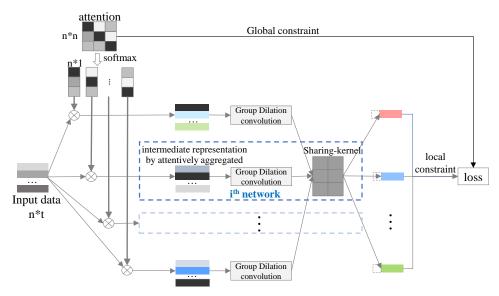
Fig. 2: Approach Overview of Shylock.

receptive field larger than the maximum lag K for accurate modeling. Existing methods often require convolution kernels with $N * K$ parameters, leading to overfitting. To balance receptive field size and parameter efficiency, we adopt group dilated convolution for univariate time-series modeling (Figure 3).

Specifically, our method employs N sets of convolution filter groups to model feature representations. To capture temporal features effectively without loss of resolution, dilated convolutions are utilized, allowing exponential expansion of the receptive field [28]. With a dilation factor of length d_l, the receptive filed reaches approximately $d_l = T^l - 1$, mitigating long time delays with only $O(LNT)$ parameters. For example, as shown in Figure 3, four variables X_1, X_2, X_3, X_4 each with eight timesteps benefit from this approach.
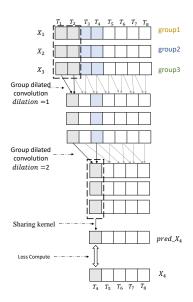

Fig. 3: Group dilated convolution.

Sharing kernel: To model variable relationships while avoiding overfitting, shared kernels are introduced. These kernels reduce interference and parameter count by focusing on local variable associations. In the i_{th} network, shared kernels model the filtered representation of the i_{th} variable after group dilated convolution, updating the attention matrix via loss computation. For variables Q, K, and V with a causal relationship $Q \rightarrow K$ (where Q represents the cause, k represents the effect, and V represents no causal relationship. So, during the causal discovery among these variables, V may introduce noise. We use $W = \{w_1, ..., w_n\}$ to represent the weight parameter of causal effect, and δ to represent the noise. They can be formalized as:

$$K = WQ + \delta, w_{i \notin q} = 0 \tag{1}$$

Therefore, the inference of causal relationships between variables does not require the participation of all data. At the same time, it's more likely that there will be a consistent set of variables depending on different variables on few-shot MTS. For all networks, the sparse model of the convolution kernel is different for the parameters of the convolution kernel only act on the association inference between local variables. So we propose sharing the kernel for the second kind of the kernel to reduce the parameters and enhance anti-interference ability.

B. DAG constraint-based global causal discovery:

Each attention-based neural network models the causal relationship for a variable as the cause and other variables as the effect. However, these networks independently focus on individual variables and cannot optimize from a global perspective to eliminate cyclic relationships. Directed acyclic graphs (DAGs) provide a strict acyclicity constraint, making them suitable for causal discovery. We formalize the causal relationship as $x_i = w_i X + \delta_i$, with attention matrix A representing causal relationships. If $W_{i,j} >$ threshold, we

define $A_{i,j} = 1$ to indicate causality. We impose a global penalty on W via $h(W)$, formulated as:

$$h(W) = tr(\sum_{k=1} A^k) \tag{2}$$

This function ensures acyclic relationships when $h(W) = 0$, and its value increases with the presence of loops. $h(W)$ also guarantees numerical stability for function and gradient evaluations. Suppose the adjacent matrix of directed acyclic garph G as A, then $a_{ij}^{(1)} = 1$ in A^1 represents there is a path from i_{th} variable to the j_{th} variable. Then $A^2 = A^1 * A$, which can be further represented as $a_{ij}^{(2)} = \sum_{p=0}^{n} a_{ip}^{(1)} a_{pj}$. If $a_{ij}^{(2)} > 0$, there is a 2-length path from $V_i \to V_j$. And so on, if $a_{ij}^{(k)}$ in $A^{(1)}$ is larger than 0, there is a k-length path from v_i to v_j. Then we can deduce that $tr(A^k) = \sum_{i=0}^{n} a_{ii}^k, k > 0$. $tr(A^k) > 0$ represents that there is a k-length path from i_{th} variable to itself. To reduce it, the coefficient $h(W)$ is applied to punish the causal loops of different lengths, denoted as: $h(W) = tr(\sum_{k=1} \beta^k A^k)$ Based on this, the whole loss of our method is:

$$\ell(\cdot) = \sum_{k=0}^{n-1} (k+1)W^k$$
$$= \sum_{k=0}^{m} \frac{1}{n} \sum_{i=0}^{n} (y_i - f(x_i|W_a^k, W_c))^2 + \alpha tr(\sum_{l=1}^{m} W_a^l) + \beta|W_c| \tag{3}$$

where we equally treat circular causal relationships of different lengths (3) can help transmit the global loss into local loss for each network. In this method, the discrete networks aiming at each variable are recombined into a global continuous optimization model, and the combination of local single-objective high-precision causal discovery and global causal graph constraint is realized.

V. EXPERIMENTS AND RESULTS

In this section, we apply Shylock to three benchmarks to find causal relationships. And compare it with two state-of-the-art works. To evaluate the efficiency and effectiveness of Shylock, we validate it on the synthetic and FMRI datasets. we answer the following research questions. **RQ1:** How is the performance of Shylock on real datasets? **RQ2:** How is the performance of Shylock on datasets with different data sizes? **RQ3:** How is the performance of Shylock on datasets with different time lags?

A. Datasets and Metrics

Datasets: In order to answer these questions, we introduce one synthetic dataset and one real dataset.

- **Synthetic Multivariate time series Datasets (Few-shot):** The synthetic datasets support customized conditions such as the num of variables, the scale of time delay, and the causal relationship. More details are described in Section V-B.
- **Real Multivariate time series Datasets (Normal):** The second benchmark is the common benchmark, functional

magnetic resonance imaging data (FMRI) [7], which is time series measuring the relationships between blood flow and different regions in the brain. We select all the 28 sub-datasets with the node_num in $\{5, 10\}$, the timesteps varies from $\{50 - 5000\}$ and the range of ds is $\{10, 12, 13, 21, 33\}$. Each variable in this dataset involves self-causation. The time delay between cause and effect is not available in FMRI.

Metrics: Performance is evaluated using standard metrics: Structural Hamming Distance (SHD), Recall also called true positive rate (TPR) ($\frac{TP}{TP+FN}$), Precision ($\frac{TP}{TP+FP}$) and F1 score ($2*\frac{Precision*Recall}{Precision+Recall}$). SHD measures the minimum edge operations (insertion, deletion, reversal) required to align the predicted graph with the true graph [21]. Precision, Recall, and F1 scores range from 0 to 1, with higher values indicating better performance. An F1 score of 1 indicates perfect model prediction. Metrics are computed by comparing the predicted attention matrix A to the formalized matrix derived from the ground truth causal graph.

B. Method for synthesizing data

Why synthetic datasets is needed? Since there are few causal datasets available in the real-world and existing synthetic methods either lack the support of sequential relation or lack support for causal relationships. So we propose a generalized method for synthesizing multivariate time series.

How is the synthetic dataset generated? To generates few-shot multivariate time series with causal relationships, we draw inspiration from [25]. First, a causal graph $\mathcal{G}$ is created as a directed acyclic graph (DAG) with n nodes, represented by an adjacency matrix M_{DAG}. The matrix is generated as a random graph similar to [25], with d total causal relationships. Each node has an average degree of d/n (in-degree and out-degree combined). Time delays for causal edges are randomly assigned within a user-defined range.

Secondly, initial time series are assigned to nodes with zero in-degree, indicating no cause. To provide flexibility for "effect" changes and better feature representation, the spline interpolation method using a three-moment equation is employed. Solving this equation yields curve functions, from which initial time series are uniformly sampled.

Thirdly, assign the degree to other nodes. A topological sort ensures nodes are processed only after their causal predecessors are generated. Time series data are computed by weighting the time lag matrix and adding noise. In the node time-series data reasoning, according to the first step in the simulation time lag matrix weighting matrix and noise generated values the simulation time sequence data of linear target node is calculated.

C. Baselines

For RQ1-3, we evaluate Shylock with two baselines NOTEARS [25] and TCDF [14]. The first work, NOTEARS is a score-based DAG structure learning method for the causal discovery method on data without temporal distribution. The second work, TCDF (Temporal Causal Discovery Framework),
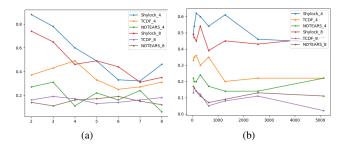
Fig. 4: F1 scores on the synthetic datasets. The chart name consists of two qualifiers. The first qualifier describes the method, the second describes node_num n.

is a constraint-based framework for causal discovery and make use of information-theoretic measures to determine dependencies between time series.

D. Implementation Details

We conducted all the experiments on a computer with Windows 10, 32 GB memory, and an Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz. For Shylock, the dilation factor d_l is initialized as 4, and the sharing kernel size as 4. Initially, the non-diagonal elements of the attention matrix A are initialized to 1. For TCDF, to make it more suitable for few-shot MTS, the kernel size is setted as 4. Other unlisted settings follow the settings of TCDF and NOTEARS.

E. Result Analysis

RQ1: In order to assess the ability of Shylock to solve the causal relationship on normal data-set, we compare it with the state-of-the art work TCDF and NOTEARS on FMRI.

TABLE I: The average results on FMRI of three methods.

model	SHD	avg_SHD	Precision	Recall	f1
Shylock	16.04	0.58	0.91	0.50	0.64
TCDF	16.79	0.75	0.59	0.56	0.57
NOTEARS	328.39	8.83	0.77	0.1	0.18

We report all the evaluation results on Table **??** in Appendix. For easy description, we report the average results of the evaluation metrics on these 28 sub-datasets in FMRI, shown as Table I. We report the evaluation results of these three methods with five repetitive times. For Shylock, each run with different initial value in attention matrix. Additionally, to fully and fairly evaluate their performance, we introduce $avg\_SHD = SHD/d$ which reflects the average distance on each causal relationship. For example, there are two datasets where $dataset_1$ involves 100 causal edges, SHD=10, and the other $dataset_2$ involves 10 causal edges, SHD=2. Though the SHD of $dataset_2$ is low, the performance of $dataset_1$ is more satisfactory.

NOTEARS performs the worst on the FMRI dataset for self-causation is not allowed. Although TCDF performs well evaluated by Recall and performs almost the same as Shylock on SHD and avg_SHD, it also has a small Precision rate, which

makes the overall F1 performance weaker than Shylock. It tends to find false positive causal relationships.

Shylock still performs the best among the three method. The F1 score can even achieves 1 in some sub-datasets reflecting it tends matain low FP. The Shylock can achieve a balanced performance of Precision and Recall, making the output results of the algorithm more confident. Even in the special case in $FMRI\_50\_200\_111\_50$, Shylock still performs the best on the $Precision, Recall, F1$.

RQ2: To address RQ2, We compare Shylock with TCDF and NoTears for causal discovery and evaluate their performance on the same two benchmarks. The first synthetic dataset is settled with node_num $n \in \{4, 8\}$, sample_num $sm \in \{40, 80, 160, 320, 640, 1280, 2560, 5120\}$ and time delay $td \in \{2, 3, 4, 5, 6, 7, 8\}$. The second dataset formed by sampling data from the FMRI data set. For each sub-dataset in FMRI, we sampled the last 40 timesteps from the original dataset.

The experimental results on synthetic dataset are shown in Figure 4(b). We can observe that under different sampling quantities, our method still achieves state-of-the-art results, which proves that Shylock not only has superior performance on few-shot datasets, but also has good results when generalized to general datasets. At the same time, it can be seen from the experimental results that with the increase of the number of samples, our method shows better performance.

Results on sampled FMRI dataset are shown in Table **??** in Section 6 and Table II. We can observe that our method Shylock still achieves state-of-the-art results, albeit with a slight drop in prediction accuracy. It proves even on the real few-shot MTS data, Shylock can work well on them.

TABLE II: The average results on sampled FMRI of three methods.

model	SHD	avg_SHD	Precision	Recall	f1
Shylock	23.61	0.82	0.58	0.53	0.55
TCDF	20.07	1.17	0.32	0.37	0.34
NOTEARS	337.61	8.94	0.78	0.13	0.22

RQ3: Figure 4(a) compares Shylock, TCDF, and NOTEARS on synthetic datasets with variables $n \in \{4, 8\}$ and time delay $td \in \{2, 3, 4, 5, 6, 7, 8\}$, while other parameters follow Section Section V-A. Each dataset includes 1 or 2 causes($d \in \{1, 2\}$). Shylock consistently achieves state-of-the-art F1 scores across nearly all cases as shown in Figure 4(a). As time lag increases, noise influences make fitting more challenging, reducing performance. Despite this, Shylock maintains superior results, demonstrating robust causal discovery capabilities.

TCDF performs worse for two reasons. Firstly, there are only small amount of data available which can not help TCDF well learn features, for its parameters is exponentially more than the amount of data itself. Secondly, there is a lack of constraints from a global perspective in TCDF. By analyzing the results of TCDF, WE find that there are a large number of causal loops in the results, most of which are the mutual

causal relationship between two nodes for each sub-networks in TCDF can only verify the causal relationship. NOTEARS performs similar in STCD_nodelay and STCD_delay for it cannot make use of the "time" features. But it can well eliminate cyclic causal relationships.

Benefiting from global causal constraints and shared parameter kernel design, Shylock enables the model to perform better on few-shot datasets with time lag.

F. Fitting ability Analysis

To assess the fitting ability of Shylock, we performed a fitting analysis. Assuming an MTS dataset with n variables, the convolution kernel was set to $n \cdot f$, where f represents the receptive field length. Most neural network-based approaches construct n sub-convolutional neural networks, each modeling a specific variable following the principle of Granger causality [1]. These sub-networks collectively extract time series features and identify causal relationships.

While using multiple sub-networks improves fitting ability, it significantly increases model complexity. For instance, TCDF [14] reduces convolution parameters compared to other methods, but each sub-network still contains at least $O(nkt)$ parameters, with the overall model having $O(nnkt)$ parameters, often approaching or exceeding the input data size for time series scenarios.

The parameter count of TCDF was validated using synthetic datasets, with 70% for training and 30% for testing. Experiments were repeated with varying sample sizes from 30 to 4000 across different random seeds, and average results were computed to minimize error accumulation. TCDF employs the Mean Squared Error (MSE) loss function, defined as:

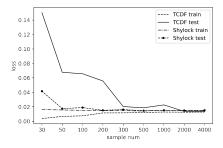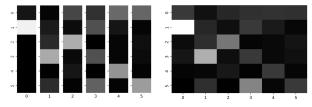$$MSEloss = \frac{1}{n}\sum_{i=0}^{n}(y_i - f(x_i|W))^2 \qquad (4)$$

Fig. 5: Comparison of fitting ability between Shylock and TCDF on datasets of different sizes.

This function effectively models the trend differences between predicted and actual values. As shown in Figure 5, when sample sizes are small, the loss gap between training and testing sets is large, indicating significant prediction deviations and overfitting during training. As sample size increases, this gap narrows, demonstrating TCDF's limited suitability for few-shot MTS data.

For fairness, NoTears was excluded as it is not a deep learning method. In contrast, Shylock shows greater stability with

(a) attention normalized by softmax

(b) Attention visualization of the dataset.

Fig. 6

small datasets, and as sample size increases, its performance approaches TCDF. This suggests Shylock effectively handles both few-shot and larger MTS datasets. TCDF's failure to model few-shot data likely stems from its large parameter space, which supports better modeling of extensive MTS data but leads to overfitting with limited samples.

G. Case study

Finally, we randomly select a dataset to illustrate the role of the attention matrix in Shylock. As shown in Figure 1, the dataset consists of $n = 6$ time series over $t = 40$ timesteps, with a time delay $td = 4$ and degree $d = 6$. The attention matrix aids in obtaining an intermediate representation of input data. Since Shylock generates sub-networks for each time series, the visualized attention matrix in Figure 6(a) reveals six sub-graphs, where brighter colors indicate higher weights. The x-axis represents potential "effect" nodes, while the y-axis represents "cause" nodes.

In Figure 6(b), the (0,1) region is notably bright, indicating a strong causal relationship between $0 \rightarrow 1$ where node 0 is the cause and node 1 is the effect. The attention matrix highlights key causal relationships in the dataset, including $5 \rightarrow 3, 3 \rightarrow 1$, and $1 \rightarrow 0$, as reflected by their bright colors and high weights. Other unrelated edges have smaller weights, demonstrating Shylock is able to accurately capture relevant causal relationships aligned with the ground truth.

VI. RELATED WORK

A range of approaches to causal discovery over time series has been proposed. They can be classified into the following classes.

Constraint-based methods [19] [24] [17] [20] are well-known two-phase procedures. They first infer the causal relationships by the conditional dependencies imprinted in the data and then search for a DAG that entails all (and only) of these dependencies. These methods do not necessarily provide complete causal information because they output (independence) equivalence classes, i.e., a set of causal structures satisfying the same conditional dependencies. One representative research, PCMCI, declared that Including more variables makes an analysis more credible regarding a causal interpretation but may lead to more side effects (e.g., leading to smaller effect sizes). It proposed to first perform a condition selection stage to remove irrelevant variables and a conditional independence test designed for highly interdependent time series. Actually, it

will introduce high noise for it has to perform a huge amount of conditional independence tests.

Score-based methods [6] [4] [5] [26], treat the causal graph as Bayesian networks. They use scoring metrics to evaluate the goodness-of-fit of the learned causal relationships and enforce the method of learning relationships toward high scores. But Score-based methods need to search high-dimension space to find the optimal result which has been demonstrated as NP-complete [2]. One of the most well-known methods, NOTEARS [25], transforms the search problem into a purely continuous optimization problem to avoid the NP-complete problem. But it still cannot model the time delay.

Deep learning-based methods [18] [23] [16] [15] [8] aim to obtain an intermediate representation that can be used to represent the characteristics of the data in a certain time window. It can be used for feature extraction of time series in a scenario with a huge amount of data. However, it cannot learn the feature well with a small amount of data and some cannot well model the time delay. In extreme cases, there may even be cases where the number of parameters in the model is exponentially larger than the actual amount of data. Then, some methods like [14] [27] use Neural Networks to improve the accuracy of results by replacing the conditional independence test and designing different rules to check its results. This kind of method still faces over-parametrization problems. For example, [14] uses multiple convolutional neural networks to model the causal relationships of each variable, and then proposes PIVM to verify the founded causal relationships. It shares the same shortcoming as Neural Network-based methods. Additionally, it's easy to see cyclic casual relationships.

VII. CONCLUSIONS

We propose Hybrid constraints-based model, **Shylock**, a new approach for causal discovery on few-shot and normal multivariate time series. Shylock is developed to better leverage fewer parameters to learn a better representation of each time series for finding local causal relationships and utilize DAG to impose global constraints to realize information sharing among sub-networks. The results of the experiment state our model is effective for causal discovery and can provide interpretability by the attention results. We also released a third-party library **Tcausal**, which has been deployed on earthDataMiner for easy use.

REFERENCES

[1] Chen, Y., Rangarajan, G., Feng, J., Ding, M.: Analyzing multiple nonlinear time series with extended granger causality. Physics letters A **324**(1), 26–35 (2004)

[2] Chickering, D.M.: Learning bayesian networks is np-complete. In: Learning from data, pp. 121–130. Springer (1996)

[3] Cui, Y., Zheng, K., Cui, D., Xie, J., Deng, L., Huang, F., Zhou, X.: Metro: a generic graph neural network framework for multivariate time series forecasting. Proceedings of the VLDB Endowment **15**(2), 224–236 (2021)

[4] Friedman, N.: The bayesian structural em algorithm. arXiv preprint arXiv:1301.7373 (2013)

[5] Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. arXiv preprint arXiv:1301.7374 (2013)

[6] Friedman, N., et al.: Learning belief networks in the presence of missing values and hidden variables. In: Icml. vol. 97, pp. 125–133. Berkeley, CA (1997)

[7] Friston, K.: Causal modelling and brain connectivity in functional magnetic resonance imaging. PLoS biology **7**(2), e1000033 (2009)

[8] Gao, Y., Shen, L., Xia, S.T.: Dag-gan: Causal structure learning with generative adversarial nets. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3320–3324. IEEE (2021)

[9] Geweke, J.: Measurement of linear dependence and feedback between multiple time series. Journal of the American statistical association **77**(378), 304–313 (1982)

[10] Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods (2001)

[11] Huang, Y., Kleindessner, M., Munishkin, A., Varshney, D., Guo, P., Wang, J.: Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. Frontiers in big Data **4**, 642182 (2021)

[12] Lee, S., Bareinboim, E.: Causal identification with matrix equations. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 9468–9479 (2021)

[13] Luo, L., Liu, W., Koprinska, I., Chen, F.: Discovering causal structures from time series data via enhanced granger causality. In: Pfahringer, B., Renz, J. (eds.) AI 2015: Advances in Artificial Intelligence - 28th Australasian Joint Conference, Canberra, ACT, Australia, November 30 - December 4, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9457, pp. 365–378. Springer (2015). https://doi.org/10.1007/978-3-319-26350-2_32, https://doi.org/10.1007/978-3-319-26350-2_32

[14] Nauta, M., Bucur, D., Seifert, C.: Causal discovery with attention-based convolutional neural networks. Mach. Learn. Knowl. Extr. **1**(1), 312–340 (2019). https://doi.org/10.3390/make1010019, https://doi.org/10.3390/make1010019

[15] Ng, I., Zhu, S., Chen, Z., Fang, Z.: A graph autoencoder approach to causal structure learning. arXiv preprint arXiv:1911.07420 (2019)

[16] Ng, I., Zhu, S., Fang, Z., Li, H., Chen, Z., Wang, J.: Masked gradient-based causal structure learning. In: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM). pp. 424–432. SIAM (2022)

[17] Ogarrio, J.M., Spirtes, P., Ramsey, J.: A hybrid causal search algorithm for latent variable models. In: Conference on probabilistic graphical models. pp. 368–379. PMLR (2016)

[18] Shang, C., Chen, J., Bi, J.: Discrete graph structure learning for forecasting multiple time series. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=WEHSlH5mOk

[19] Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)

[20] Triantafillou, S., Tsamardinos, I.: Score-based vs constraint-based causal learning in the presence of confounders. In: Cfa@ uai. pp. 59–67 (2016)

[21] Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine learning **65**(1), 31–78 (2006)

[22] Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C.: Connecting the dots: Multivariate time series forecasting with graph neural networks. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 753–763 (2020)

[23] Yu, Y., Chen, J., Gao, T., Yu, M.: Dag-gnn: Dag structure learning with graph neural networks. In: International Conference on Machine Learning. pp. 7154–7163. PMLR (2019)

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

[25] Zheng, X., Aragam, B., Ravikumar, P., Xing, E.P.: Dags with no tears: Continuous optimization for structure learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9492–9503. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)

[26] Zheng, X., Aragam, B., Ravikumar, P.K., Xing, E.P.: Dags with no tears: Continuous optimization for structure learning. Advances in Neural Information Processing Systems **31** (2018)

[27] Zheng, X., Dan, C., Aragam, B., Ravikumar, P., Xing, E.: Learning sparse nonparametric dags. In: International Conference on Artificial Intelligence and Statistics. pp. 3414–3425. PMLR (2020)

[28] Zhou, H., Zhu, Y., Wang, Q., Xu, J., Li, G., Chen, D., Dong, Y., Zhang, H.: Multi-scale dilated convolution neural network for image artifact correction of limited-angle tomography. IEEE Access **8**, 1567–1576 (2020). https://doi.org/10.1109/ACCESS.2019.2962071