

An Interpretable Temporal Convolutional Framework for Granger Causality Analysis

Abstract— Most existing parametric approaches for detecting linear or nonlinear Granger causality (GC) face challenges in estimating appropriate time delays, a critical factor for accurate GC detection. This issue becomes particularly pronounced in nonlinear complex systems, which are often opaque and consist of numerous components or variables. In this paper, we propose a novel temporal convolutional network (TCN)-based end-to-end GC detection approach called the Interpretable Temporal Convolutional Framework (ITCF). Unlike conventional deep learning models, which act like a “black box” and are difficult to analyse the interactions between variables, the proposed ITCF is able to detect both linear and nonlinear GC and automatically estimate time delay during the multivariate time series prediction. Specifically, GC is obtained by employing the Least Absolute Shrinkage and Selection Operator (Lasso) regression during the prediction of multivariate time series using TCN. Then, time delays can be estimated by interpreting the TCN kernels. We propose a convolutional Hierarchical Group Lasso (cHGL), a hierarchical regularisation approach to effectively utilise temporal information within each TCN channel for enhanced GC detection. Additionally, as far as we are concerned, this paper is the first to integrate the Iterative Soft-Thresholding Algorithm into the backpropagation of TCN to optimise the proposed cHGL, which enabling causal channel selection and inducing sparsity within each TCN channel to remove redundant temporal information, ultimately creating an end-to-end GC detection framework. The testing results of four experiments, involving two simulations and two real data, demonstrate that the proposed ITCF, in comparison with state-of-the-art, offers a more reliable estimation of GC relationships in complex systems featuring intricate dynamics, limited data lengths, or numerous variables.

Index Terms—Granger causality, Interpretable deep learning, Iterative Soft-Thresholding Algorithm (ISTA), Least Absolute Shrinkage and Selection Operator (Lasso), temporal convolutional network (TCN).

I. INTRODUCTION

COMPLEX systems feature a large number of measurable variables interacting with each other on multiple levels [1]. In real-world scenarios, opacity and nonlinearity widely exist in complex systems, such as the price relationships among international stock markets in financial systems [2], [3], brain functional connectivity in neuro systems [4], [5], the interactions of atmospheric, oceanic, and terrestrial components in climate systems [6], [7], etc. To discover the underlying interactions between different measurable components, the Granger Causality (GC) test was proposed by assessing whether the historical information of one time series

helps to forecast another [8]. Over time, the methodological research in GC has broadened its scope from detecting linear interactions to nonlinear interactions between variables [9].

Generally, there are two types of GC detection methods: parametric and nonparametric. The nonparametric approaches, such as transfer entropy [10], [11], spectrum analysis [12], [13], and kernel methods [14], [15], etc., analyse the observed data directly without relying on models, thereby offering greater flexibility. However, long time series data is usually required to mitigate the effects of noise for reliable estimation in nonparametric methods [16]. Furthermore, they are susceptible to curse of dimensionality with high-dimensional multivariate datasets [17]. In contrast, the parametric methods are typically more efficient in terms of data volume requirements as they extract causality information based on model structure assumptions and estimations [16]. Many conventional parametric approaches assume the interactions are linear, where the vector autoregressive (VAR) model is widely used in multivariate scenario [18]. Least absolute shrinkage and selection operator (Lasso) is commonly employed in model structure estimation in order to induce sparsity into the regression process [19]. In GC detection with VAR models, parameters are allocated for all potential causal time series. The parameters of each time series are regarded as a group. Then the Group Lasso (GL) was used to encourage parameters in the same group to be set to zero simultaneously [20], [21]. However, this approach does not consider the potential temporal relationships within each group. To tackle this problem, the Hierarchical Group Lasso (HGL) was developed by grouping parameters of each time delay hierarchically into nested groups, inducing sparsity pattern in lag parameters [22]. Recently, Lasso regression was implemented to calculate the transfer entropy of Gaussian processes, which is equivalent to GC in this context [23]. Furthermore, group Lasso was employed in a comparable manner to identify a VAR model for causality estimation [24]. However, nonlinearity widely exists in real world scenario, and linear methods normally fail to detect nonlinear GC between variables [16].

For the nonlinear parametric GC test, the regressor library-based methods have been developed and prevailed in the past few years [25]. In the regressor library, all potential causal variables are converted into different nonlinear candidate terms, including polynomials [26], [27], radial basis functions [28], [29], etc., and enumerated within a specified maximum time delay. Then the causal variables and appropriate time delay can

be selected with forward regression with orthogonal least squares (FROLS). However, over-specifying the possible maximum time delay will exponentially increase the number of candidate terms, leading to prolonged computational time and a higher risk of overfitting. Furthermore, procedures of model structure estimation become more complicated and time consuming when dealing with complex systems that are highly nonlinear and contain a large number of variables [25]. Recently, nonparametric methods have been incorporated into parametric nonlinear GC tests to reduce the dimension of candidate regressors in the multivariate scenario. In [30], the state space reconstruction was utilised to transform the original measurement space into non-uniform embeddings (NUE). Then the Lasso regression has been employed in the reproducing kernel Hilbert spaces (RKHS) of the NUE. Finally, kernels that have been selected by Lasso can be retraced back to the original measurement space to identify the causal components. However, the selection of appropriate time delay still seriously affects the accuracy of GC detection [30].

With the growth of computational power, neural networks (NN) have become more popular in time series modelling for highly nonlinear and complex systems because of their prominent universal approximation ability and less reliance on prior knowledge [31], [32]. In time series modelling, one major objective of neural network architecture design is to minimise the number of parameters, thereby mitigating the risk of overfitting [33]. When employing the vanilla multilayer perceptron (MLP) for time series modelling, weights are assigned to each time step to extract temporal information, which makes it more prone to overfitting when the length of the input time series increases [34]. To tackle the problem of overfitting, Recurrent neural networks (RNN), such as the vanilla RNN, long-short-term memory (LSTM) and gated recurrent unit (GRU), are developed with weights sharing strategy [35], [36], [37]. In RNNs, each time step of the input time series shares the same set of weights. Therefore, the parameter dimension in RNN is irrelevant to the length of the input time series [38]. In training, the RNN weights are updated by backpropagation through time (BPTT) to learn the temporal information [39]. Similarly, the weights-sharing strategy is adopted in temporal convolution networks (TCN), where the kernel weights slide sparsely across the input time series to extract the temporal information [40], [41]. Although these methods have demonstrated promising forecasting performance, they provide limited interpretability of the interactions between different time series. Granger causality is inherently about explaining whether and how the past values of one variable contribute to the future values of another. While deep learning models like Recurrent Neural Networks (RNNs) and Temporal Convolutional Networks (TCNs) can model such dependencies, they do not provide explicit insights into which variables are influencing the predictions or through which time lags. This opacity makes it challenging to identify which variables are causally significant, undermining the primary objective of causality analysis.

Recently, there has been a surge in research focusing on neural network-based GC tests, which can be divided into two

types generally. The first type of approach evaluates whether a specific input time series contributes to output prediction by diminishing the prediction error. In [42], the NN-GC framework was proposed where the GC is tested by calculating the difference between prediction errors of MLP with and without the past information of certain input variables using the NUE strategy. Since the vanilla MLP is only capable of extracting causal information from the entire input time series rather than the time series within an appropriate time delay, the input time series length seriously affects the accuracy of GC tests. Therefore, the NN-GC framework selects the appropriate time delay and truncates the input time series with the NUE strategy. However, the time delay selection still affects the computational time and preserves the risk of overfitting of NUE-based MLP. To tackle this problem, the RNN-GC [38] and seq2seq-LSTM Granger Causality (SL-GC) [43] frameworks have been proposed. Since the parameter dimension in RNNs is irrelevant to the length of the input time series, the RNN-GC and SL-GC are less sensitive to the time delay selection of input time series, resulting in a higher accuracy of causality detection [38], [43]. However, the NNs are required to be trained multiple times to evaluate the contribution of each input in this type of method, making them relatively more complicated than the following end-to-end approaches.

The second type of neural network-based GC detection method is based on sparse regression where the causal inputs are detected in an end-to-end manner. In [44], the Lasso penalty has been incorporated into the training of two types of NNs including the component-wise MLP (cMLP) and component-wise LSTM (cLSTM). The cMLP and cLSTM were proposed to disentangle the contribution of different inputs to the outputs. In the training of cMLP, the causal inputs and corresponding time delay are selected together by applying HGL which induces both temporal and component-wise sparsity. Therefore, it offers a more efficient way for nonlinear GC detection compared to regressor library-based methods [44]. However, the fully connected structure in hidden layers of cMLP still undermines its generalisation performance. For the cLSTM, only the causal inputs can be selected while the appropriate time delay information remains unknown. In the training, only the GL can be applied on cLSTM due to the weights-sharing structure of RNNs. It is not feasible to examine the individual contribution of a parameter at a specific time step while selecting the causal inputs. Therefore, it leads to a deterioration in performance in two simulation datasets, despite RNNs typically outperforming MLPs in time series forecasting [44]. On the other hand, attention-based methods have been introduced in TCN. In [45], the temporal Causal Discovery Framework (TCDF) was proposed, where each input channel of the TCN is assigned a trainable attention score. Causal inputs are then detected by applying a threshold to these attention scores. Then the corresponding time delay can be discovered by weight interpretation due to the dilation structure of the TCN. However, the causal input selection only relies on channel-wise attention while the temporal information is not considered in the process of GC detection. To tackle this problem, temporal

attention has recently been incorporated into TCN to reduce the redundant temporal information in causal inputs [46]. In [46], the linear causal inputs are selected based on Lasso regression in the data preprocessing rather than through the end-to-end training of the TCN.

Current methods for Granger causality detection face several limitations, particularly when applied to highly nonlinear complex systems. The traditional approaches, whether parametric or nonparametric, often struggle with challenges such as accurately estimating time delays, dealing with high-dimensional data, and high degree of nonlinearity. Additionally, neural network-based Granger causality approaches often lack interpretability of input time delays and are sensitive to the selection of time delays, often including redundant temporal information, which reduces their reliability in complex scenarios.

In this work, we proposed an end-to-end GC detection framework based on sparse regression of TCN, namely the Interpretable Temporal Convolutional Framework (ITCF), to detect both linear and nonlinear GC relationships in multivariate time series. Within the ITCF, we introduced the Convolutional Hierarchical Group Lasso (cHGL) in the training of the TCN for selecting causal inputs. Furthermore, to the best of our knowledge, we are the first to integrate the Iterative Soft-Thresholding Algorithm (ISTA) into the backpropagation of TCN to optimise the proposed cHGL, which enables causal channel selection and induces sparsity within each TCN channel to remove redundant temporal information, ultimately creating an end-to-end GC detection framework. The contribution of this work can be summarized as follows:

- The Interpretable Temporal Convolutional Framework (ITCF) is capable of detecting GC and automatically selecting the corresponding time delay of a complex system which is highly nonlinear and contains a large number of variables. This framework provides an interpretable deep learning solution to the challenge encountered in conventional regressor library-based approach [25].
- The proposed convolutional Hierarchical Group Lasso (cHGL) is able to utilise the temporal information within each input time series to facilitate the selection of causal inputs. Consequently, it mitigates the limitations present in the TCDF [45], resulting in improved performance.
- To optimise the cHGL, we integrate the Iterative Soft-Thresholding Algorithm (ISTA) into the backpropagation process of TCN. This innovation facilitates the end-to-end GC detection, offering greater efficiency compared to approaches that employ Lasso in data preprocessing [46], [47].

II. NEURAL NETWORK-BASED GRANGER CAUSALITY

Granger causality examines whether the historical values of one time series, X , are useful in forecasting the future values of another time series, Y , beyond what the past of Y can predict alone [8]. The Granger causality of X to Y , $G_{X \rightarrow Y}$, is defined by

$$G_{X \rightarrow Y} = \ln \frac{\text{var}(Y | Y^-)}{\text{var}(Y | Y^-, X^-)}, \quad (1)$$

where X^- and Y^- denote the past information of time series X and Y , respectively.

In [44], Granger causality was extended from univariate to multivariate time series analysis. Consider a multivariate time series $\mathbf{Y}$ comprising N variables, each with T samples, denoted as $\{\mathbf{Y}^t\}_{t \in [1, T]} = \{(y_1^t, y_2^t, \dots, y_j^t, \dots, y_N^t)\}_{t \in [1, T]}$. If Granger causality exists between variables in $\mathbf{Y}$, it can be described as

$$y_n^t = g_n(y_1^-, \dots, y_j^-, \dots, y_N^-) + e_t^n, \quad 1 \leq n \leq N, \quad (2)$$

where $y_j^- = (y_j^{t-1}, y_j^{t-2}, \dots, y_j^{t-K_j})$ denotes the historical values of the time series y_j within the maximum time delay K_j , and e_t is the additive zero mean noise. The functions $g_n(\cdot)$ specify how the past information of the time series $[y_1, \dots, y_N]$ is mapped to the series y_n . In the context of neural network-based Granger causality test, $g_n(\cdot)$ are represented by neural networks.

The null hypothesis (H_0), stating that the time series y_j is Granger non-causal for the time series y_n for all $t \in [1, T]$, is defined as

$$g_n(y_1^-, \dots, y_j^-, \dots, y_N^-) = g_n(y_1^-, \dots, y_{j-1}^-, y_{j+1}^-, \dots, y_N^-). \quad (3)$$

Detecting non-causal time series pair (y_n, y_j) in the multivariate time series $\mathbf{Y}$ is achieved through the application of the Lasso penalty in regression, including Group Lasso [48] and Hierarchical Lasso [49]. In [44], the Hierarchical Lasso outperforms other Lasso penalties since it takes into account the lag of each Granger causal pair. Therefore, this research adopts the Hierarchical Lasso, with further details described in Section III.

III. METHODOLOGY

In this work, we introduce an Interpretable Temporal Convolutional Framework (ITCF) for the estimation of Granger causality in multivariate time series. The ITCF comprises two main components: (a) Granger causality detection using a new proposed sparse regression based TCN (sTCN) and (b) maximum time delay estimation by kernel interpretation. The sTCN is designed to detect Granger causality by applying a new convolutional Hierarchical Group Lasso (cHGL) during the prediction of multivariate time series. To optimise cHGL based on sTCN, we incorporate the iterative soft-thresholding algorithm (ISTA) into the backpropagation process. The following sections begin by introducing the architecture of multivariate time series prediction using the sTCN, followed by a detailed description of the mechanism of cHGL and the implementation of the ISTA during the backpropagation. Lastly, we describe the algorithm for estimating the maximum time delay.

A. Architecture of Multivariate Time Series Prediction

In this research, the sparse regression based TCN (sTCN) is adopted as the model $g_n(\cdot)$ which is responsible for mapping the past information of the time series $[y_1, \dots, y_N]$ to predict the series y_n in (2). To detect Granger causality in the multivariate time series $\mathbf{Y} = [y_1, \dots, y_N]$, each $y_n, \forall y_n \in \mathbf{Y}$ is sequentially selected as the target for prediction. Then all time series in $\mathbf{Y}$ are considered as potential causal inputs and shifted by one time step

into the past. These lagged time series are then fed into separate channels of the sTCN for analysis.

The topology of the sTCN is illustrated in Fig. 1. The sTCN is constructed using multiple one-dimensional dilated convolutional (Conv1d) layers. In Fig. 1, the kernel size of each Conv1d layer is set as 2. For the L^{th} Conv1d layer, the dilation size D is given by

$$D = 2^{L-1}. \quad (4)$$

The receptive field (R) of each element in the hidden feature map, derived by the L^{th} Conv1d layer is calculated as:

$$R = 1 + \sum_{i=1}^L 2^{i-1}. \quad (5)$$

To predict y_n , each time series in $\mathbf{Y}$ is considered as a potential causal input and is then fed into separate channels of the sTCN. In each time series of $\mathbf{Y}$, there is a total of T samples. The input time series $\mathbf{Y}^{1:T-1} = [\mathbf{Y}^1, \mathbf{Y}^2, \dots, \mathbf{Y}^{T-1}]$ is delayed by one time step compared to the output $y_n^{2:T} = (y_n^2, y_n^3, \dots, y_n^T)$.

When the time series y_j is fed into channel j of the sTCN with a kernel size of k , the kernel weights in the L^{th} Conv1d layer can be represented as $f_j^L = [f_{(j,0)}^L, \dots, f_{(j,k-1)}^L]$. The output feature map of the first Conv1d layer in channel j , denoted as H_j^1 , is given by

$$H_j^1(t) = (y_j * f_j^1)_D(t) = \sum_{i=0}^{k-1} f_{(j,i)}^1 \cdot y_j^{t-D \cdot i} + b_j^1, \quad (6)$$

where $H_j^1(t)$ represents the t th element in the feature map, and $t - D \cdot i$ accounts for the direction of the past with the dilation of D in the input time series y_j . b_j^1 represents the bias in the first layer of channel j . The output H_j^1 is then activated by a parametric rectified linear unit element-wise activation ($PReLU$) function. The subsequent output of the L^{th} hidden Conv1d layers is given by a similar form, also with $PReLU$ activation functions, and can be written as:

$$H_j^L(t) = (H_j^{L-1} * f_j^L)_D(t) = \sum_{i=0}^{k-1} [f_{(j,i)}^L \cdot H_j^{L-1}(t - D \cdot i)] + b_j^L. \quad (7)$$

In each Conv1d layer, the input applies left zero padding with a padding size equals to the dilation size D . The padding is then trimmed to ensure that the length of the input matches the length of the output of the Conv1d layers. In the final output layer, there is a 1×1 convolutional layer acting as a channel-wise fully connected layer where each element in all channel outputs is combined to predict the corresponding element in the estimated output, denoted as $\hat{y}_n$.

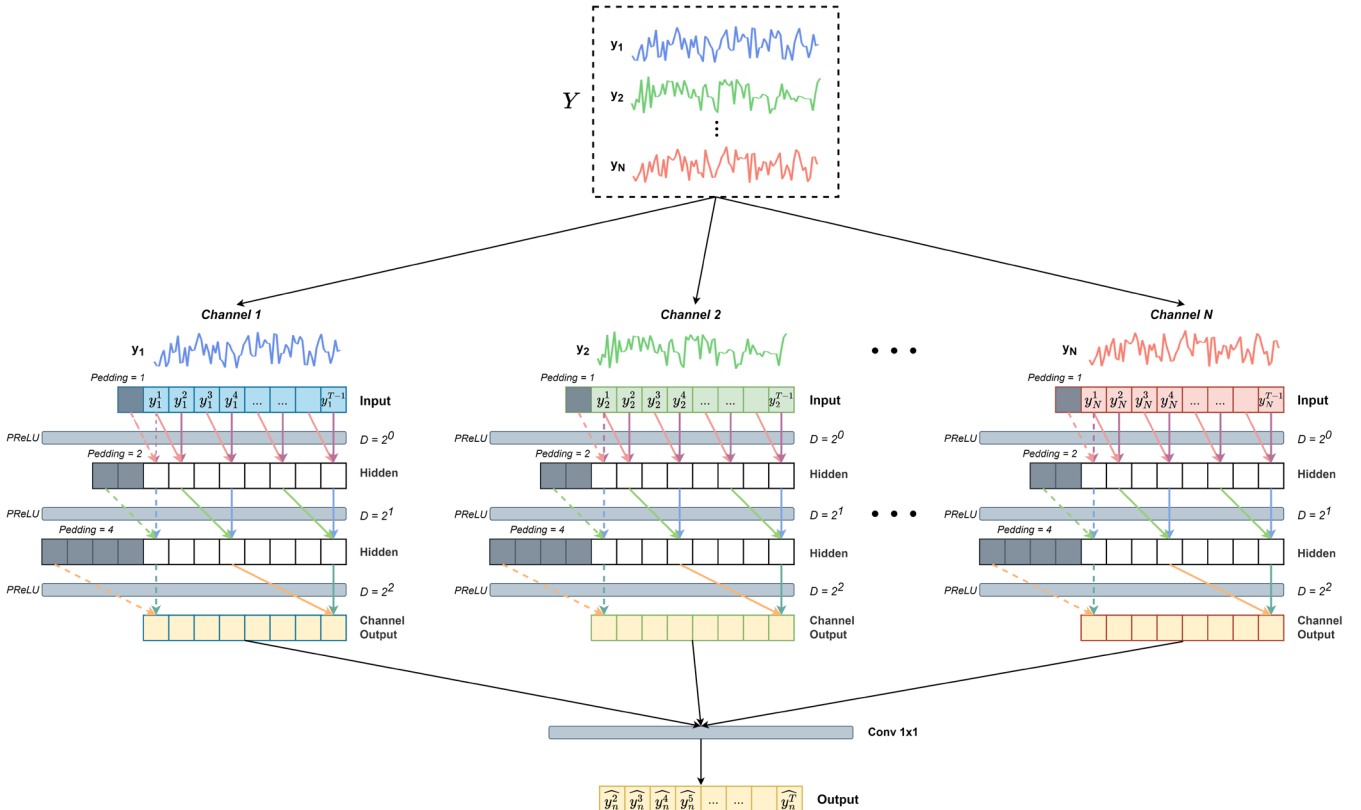


Fig. 1. The illustration of the topological structure of sTCN. The sTCN with N channels are expanded along channels. To detect Granger causality in a multivariate time series $\mathbf{Y} = [y_1, \dots, y_N]$, each time series $\forall y_n \in \mathbf{Y}$ is sequentially selected as the target for prediction. Then all time series in $\mathbf{Y}$ are considered as potential causal inputs and shifted by one time step to the past. These lagged time series are then individually fed into separate channels of the sTCN. To identify the non-causal inputs, a Hierarchical Lasso penalty is applied to the first Conv1d layer of the sTCN. Finally, the maximum time delay for each causal input can be determined by interpreting the kernel weights.

B. Causality Detection with Convolutional Hierarchical Group Lasso

According to (6), when all k kernel weights in the first Conv1d layer of channel j are zeros, y_j^- does not influence all subsequent hidden layers. This implies that the time series y_j is Granger non-causal for the time series y_n for all $t \in [1, T]$. To identify the Granger causal inputs, we propose a novel convolutional Hierarchical Group Lasso (cHGL) in the first Conv1d layer of the sTCN when predicting $\forall y_n \in \mathbf{Y}$. The objective function of the sTCN with cHGL is given by

$$\hat{\mathbf{f}} = \underset{\mathbf{f}}{\operatorname{argmin}} \left\{ \frac{1}{T-1} \sum_{t=2}^T (y_n^t - g_n(\mathbf{Y}^{t-1}))^2 + \lambda \sum_{j=1}^N \|\mathcal{G}(f_j^1)\|_1 \right\}, \quad (8)$$

where $\|\cdot\|_1$ is the L_1 norm of the channel-wise groups. λ is the cHGL hyperparameter that controls the sparsity of the kernel weights in the first Conv1d layer. The hierarchical structured convolutional kernels $\mathcal{G}(f_j^1)$ in the first layer of each channel are defined as

parameters in the same group to be set to zero simultaneously where parameters of each time series are regarded as a group [20], [21]. However, this approach does not consider the potential temporal relationships within each group. To address this limitation, the HGL was developed by grouping parameters of each time delay hierarchically into nested groups, inducing sparsity pattern in lag parameters [22]. In [44], the HGL was applied to component-wise multilayer perceptron (cMLP), where weights are allocated for each time step in the MLP. But for recurrent neural networks (RNN), weights are shared across each time step, making it impossible to inspect the individual contribution of a parameter at a specific time step in RNN. Therefore, only GL can be applied to RNN, leading to a deterioration in performance, despite RNNs typically outperforming MLPs in time series forecasting [44].

Although the kernels in the first Conv1d layer of sTCN slide sparsely across the input time series rather than being allocated for each time step as the weights in MLP, it is still possible to investigate the temporal information of each time series contained in these kernel weights. To achieve this, we hierarchically allocate the kernels in the first layer of sTCN into nested groups when implementing the cHGL. A detailed illustration of cHGL can be found in Fig. 2, which depicts the hierarchical nested groups of kernel weights for each channel enclosed in coloured boxes. These nested groups are then

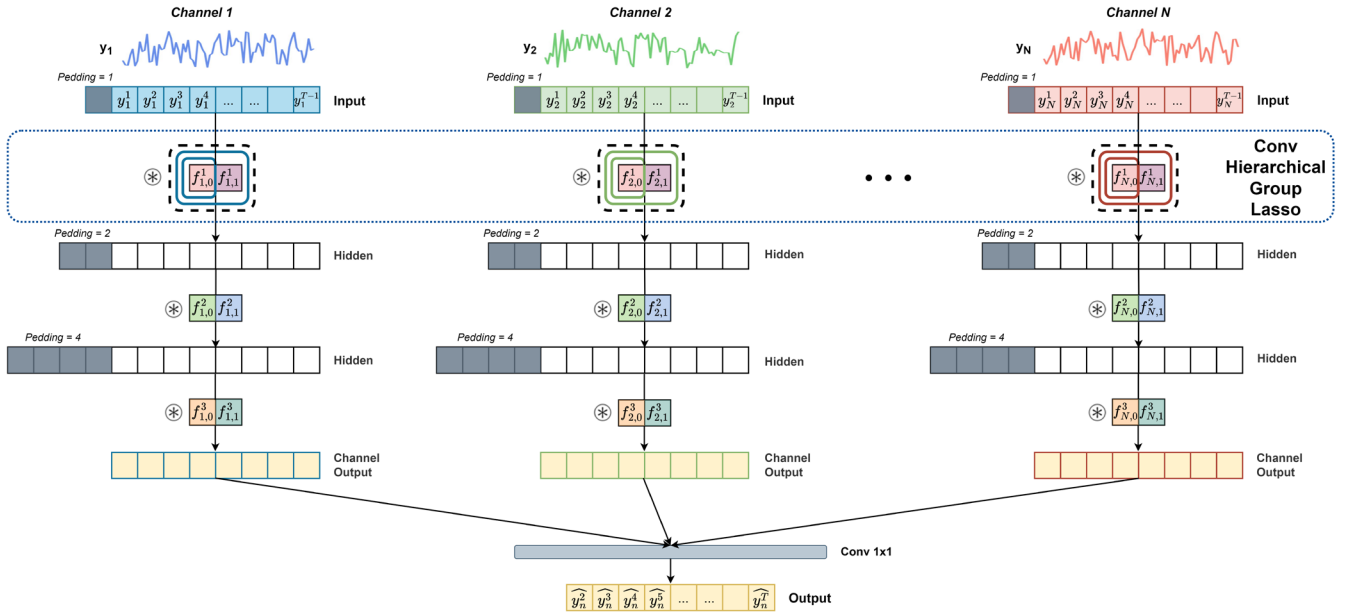


Fig. 2. An illustration of the proposed convolutional Hierarchical Group Lasso (cHGL) implemented on the sTCN with a kernel size of 2. The coloured boxes within the first Conv1d layer represent the hierarchical nested groups of kernel weights. The black dash-line boxes represent the channel-wise groups.

$$\mathcal{G}(f_j^1) = \sum_{i=0}^{k-1} \|f_{(j,i)}^1\|_F \quad (9)$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm and $f_{(j,i)}^1 = [f_{(j,0)}^1, \dots, f_{(j,i)}^1]$.

The proposed cHGL is motivated by the Hierarchical Group Lasso (HGL) and is specifically designed for the regularisation of convolutional kernels. In conventional vector autoregression (VAR) modelling, the Group Lasso (GL) is used to encourage

incorporated into channel-wise groups, represented as black dash-line boxes.

Since each kernel in the first Conv1d layer of sTCN gathers information from multiple times steps, cHGL does not directly induce the temporal sparsity, which would indicate the exact time delay of each causal input, as observed in HGL [22], [44]. Instead, cHGL uses hierarchical temporal information within each channel-wise group to facilitate the selection of causal inputs for the sTCN, thus mitigating the performance deterioration observed in GL-based RNN. The time delay

information for each causal input can still be determined by interpreting the kernel weights, thanks to the dilation structure of sTCN. More details on the process of time delay estimation will be introduced in Section D.

C. Proximal Gradient Descent

The proximal gradient descent approach was designed to optimise non-smooth objective functions by applying the proximal operator on the regularised parameters [50]. In the Lasso problem, the iterative soft-thresholding algorithm (ISTA) [51], a first-order proximal gradient descent method, is commonly used to optimise non-smooth objective function. In [52], the ISTA has been applied on the MLP to optimise the Lasso problem. In [44], ISTA was effectively incorporated into the backpropagation of MLP and RNN models to induce sparsity in the parameters of the first layer. On the other hand, the HGL problem has been effectively optimised using ISTA, as demonstrated in [53]. Therefore, drawing inspiration from [44], [53], we adapted the ISTA and incorporated it into the backpropagation process of hierarchical convolutional kernels to optimise the cHGL-based sTCN.

Unlike other existing neural network-based ISTAs including the cLSTM that are used solely for selecting causal inputs, the ISTA in this manuscript is designed to both select causal channels and induce sparsity within each TCN channel to remove redundant temporal information. To optimise the non-smooth objective function as defined in (8), the proximal gradient descent with ISTA can be derived as follows. Let the smooth part of the objective function be $F(f) = \frac{1}{T-1} \sum_{t=2}^T (y_n^t - g_n(Y^{t-1}))^2$, then (8) can be written as:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \left\{ F(f) + \lambda \sum_{j=1}^N \|\mathcal{G}(f_j^1)\|_1 \right\}. \quad (10)$$

At $\forall$ gradient decent step k , the Taylor Series quadratic approximation of the $F(f)$ around f^k is given by:

$$F(f) \approx F(f^k) + \nabla F(f^k)^T (f - f^k) + \frac{1}{2} (f - f^k)^T \nabla^2 F(f^k) (f - f^k), \quad (11)$$

where $\nabla F(f^k)$ can be calculated through backpropagation. To avoid the complicated computation process of the Hessian matrix $\nabla^2 F(f^k)$, the $\nabla^2 F(f^k)$ is replaced by $\frac{1}{\eta} I$, where η is the tuneable learning rate and I is the identity matrix. As the $F(f^k)$ is a constant, which does not affect the optimisation process and can be removed. Then the objective function (10) can be written as:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \left\{ \nabla F(f^k)^T (f - f^k) + \frac{1}{2\eta} \|f - f^k\|_2^2 + \lambda \sum_{j=1}^N \|\mathcal{G}(f_j^1)\|_1 \right\}. \quad (12)$$

By forming the perfect square trinomial, (12) can be written as:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \left\{ \frac{1}{2\eta} \|f - f^k + \eta \nabla F(f^k)\|_2^2 - \frac{\eta}{2} \|\nabla F(f^k)\|_2^2 + \lambda \sum_{j=1}^N \|\mathcal{G}(f_j^1)\|_1 \right\}, \quad (13)$$

Where the $-\frac{\eta}{2} \|\nabla F(f^k)\|_2^2$ is also a constant term without containing the variable f , therefore it can be removed in the optimisation. Then the objective function can be written as:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \left\{ \frac{1}{2\eta} \|f - (f^k - \eta \nabla F(f^k))\|_2^2 + \lambda \sum_{j=1}^N \|\mathcal{G}(f_j^1)\|_1 \right\}, \quad (14)$$

Let the variable f take gradient descent on the smooth part of the objective function be $Z_f^{k+1} = f^k - \eta \nabla F(f^k)$, then the object function can be written as:

$$\hat{f} = \underset{f}{\operatorname{argmin}} \left\{ \frac{1}{2\eta} \|f - Z_f^{k+1}\|_2^2 + \lambda \sum_{j=1}^N \|\mathcal{G}(f_j^1)\|_1 \right\}. \quad (15)$$

When reaching any optimum, the gradient of the variable become zero. For the kernel in the first layer in the channel j , the gradient with respect to the objective function (15) can be written as:

$$\hat{f}'|_{f_j^1} = \left(\frac{1}{2\eta} \|f - Z_f^{k+1}\|_2^2 + \lambda \sum_{j=1}^N \|\mathcal{G}(f_j^1)\|_1 \right)' = 0 \quad (16)$$

If the the number of kernels in channel j equals 2. Then the channel-wise root group $\mathbf{G}(f_j^1) = (\{f_{j,1}^1, f_{j,2}^1\})$ contains the children group $(\{f_{j,1}^1\})$ and $(\{f_{j,2}^1\})$. The hierarchical structure can be illustrated in Fig. 3 according to [49].

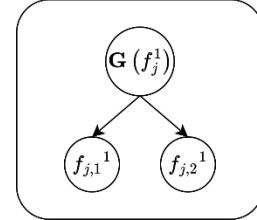


Fig. 3. The hierarchical group structure of the convolutional kernels in the first layer of channel j .

The children groups only exist when the root group exists [49]. Then, for each element in the channel-wise root group $\mathbf{G}(f_j^1)$, the soft thresholding function can be derived as follows:

$$\hat{f}'|_{f_{j,n}^1} = \left(\frac{1}{2\eta} \|f - Z_f^{k+1}\|_2^2 + \lambda \|\mathbf{G}(f_j^1)\|_F \right)' = 0 \quad (17)$$

$$\Rightarrow \frac{1}{\eta} (f_{j,n}^{1,k+1} - Z_{f_{j,n}^1}^{k+1}) + \lambda \frac{f_{j,n}^1}{\|\mathbf{G}(f_j^1)\|_F} \Big|_{Z_{f_{j,n}^1}^{k+1}} = 0 \quad (18)$$

$$\Rightarrow \frac{1}{\eta} (f_{j,n}^{1,k+1} - Z_{f_{j,n}^1}^{k+1}) + \lambda \frac{Z_{f_{j,n}^1}^{k+1}}{\|\mathbf{G}(f_j^1)\|_F} = 0 \quad (19)$$

$$\Rightarrow f_{j,n}^{1,k+1} = \left(1 - \frac{\lambda \eta}{\|\mathbf{G}(f_j^1)\|_F} \right) Z_{f_{j,n}^1}^{k+1} \quad (20)$$

This holds only when all $f_{j,n}^{1,k+1} > 0$ or in other words $\|\mathbf{G}(f_j^1)\|_F > \lambda \eta$, otherwise all $f_{j,n}^{1,k+1} = 0$. Where n is the element index in channel-wise root group $\mathbf{G}(f_j^1)$, here $n = 1$ or 2 . $Z_{f_{j,n}^1}^{k+1}$ represents the value of the variable $f_{j,n}^1$ after taking one step of gradient descent on the smooth part of the objective function.

If $\|\mathbf{G}(f_j^1)\|_F > \lambda\eta$, then the children groups exist. For $\forall (\{f_{j,n}^1\}) \supset \mathbf{G}(f_j^1)$, the soft thresholding function can be derived as follows:

$$\hat{f}'|_{f_{j,n}^1} = \left(\frac{1}{2\eta} \|f - Z_{f_j^{k+1}}\|_2^2 + \lambda \|f_{j,n}^1\|_F \right)' = 0 \quad (21)$$

$$\Rightarrow \frac{1}{\eta} \left(f_{j,n}^{1,k+1} - Z_{f_{j,n}^1}^{k+1} \right) + \lambda \frac{f_{j,n}^1}{\|f_{j,n}^1\|_F} = 0 \quad (22)$$

$$\Rightarrow \frac{1}{\eta} \left(f_{j,n}^{1,k+1} - Z_{f_{j,n}^1}^{k+1} \right) + \lambda \frac{Z_{f_{j,n}^1}^{k+1}}{\|f_{j,n}^1\|_2} = 0 \quad (23)$$

$$\Rightarrow f_{j,n}^{1,k+1} = \left(1 - \frac{\lambda\eta}{\|f_{j,n}^1\|_2} \right) Z_{f_{j,n}^1}^{k+1} \quad (24)$$

This holds only when $\|f_{j,n}^1\|_F > \lambda\eta$, otherwise $f_{j,n}^{1,k+1} = 0$. Therefore, it creates the sparsity in the first layer within each existing channel. The sparsity is illustrated in Fig. 4.

The elimination of the redundant kernel removes its contribution to the next iteration's channel-wise root group selection. Therefore, it helps remove redundant temporal information and improves Granger causality detection. In contrast, in [45], kernels in each channel-wise group are selected together, and redundant temporal information is considered during the selection process.

Therefore, the general ISTA for optimizing the cHGL can be written as follows: the proximal gradient descent for the convolutional kernel i in the first layer of channel j at iteration s is defined as

$$\mathbf{f}_{(j,i)}^{1,s} = \text{Prox}_{\lambda\eta}^{(s)} \left(\mathbf{f}_{(j,i)}^{1,(s-1)} - \eta \nabla_{\mathbf{f}_{(j,i)}^1} \mathcal{L}(\mathbf{f}^{(s-1)}) \right), \quad \forall \mathbf{f}^1 \in \mathbb{R}^{N \times k}, \quad (25)$$

where $\mathcal{L}(\mathbf{f}^{(s)})$ represents the unregularised loss function of the sTCN, defined as

$$\mathcal{L}(\mathbf{f}^{(s)}) = \frac{1}{T-1} \sum_{t=2}^T (y_n^t - g_n(\mathbf{Y}^{t-1}))^2. \quad (26)$$

η is the learning rate, and $\nabla_{\mathbf{f}_{(j,i)}^1}$ is the gradient of $\mathbf{f}_{(j,i)}^1$ derived from $\mathcal{L}(\mathbf{f}^{(s-1)})$ during backpropagation to the first layer. For

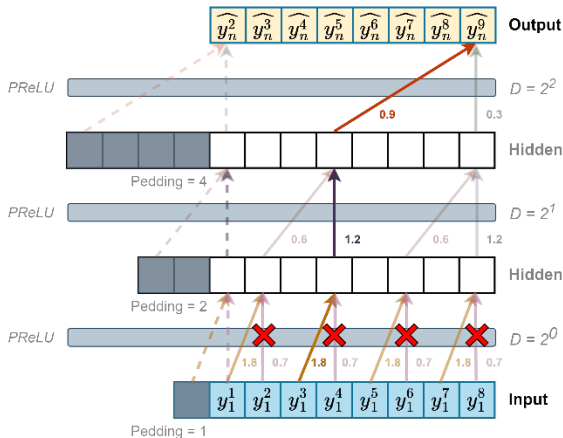


Fig. 4. The illustration of the elimination of the redundant kernels after one iteration of ISTA optimising the cHGL. The red cross represents the elimination of the kernel.

Algorithm 1: Optimising the cHGL with Iterative Soft-Thresholding Algorithm (ISTA)

Let L be the number of Conv1d layers in sTCN

Let k be the kernel size in each Conv1d layer.

Let $\{\mathbf{f}_{(j,i)}^1\}$ be the kernel weights in the first layer of channel j of the sTCN from the first weight to index i , $\{\mathbf{f}_{(j,i)}^1\} = [f_{(j,0)}^1, \dots, f_{(j,i)}^1]$, $\forall j \in [0, N], \forall i \in [1, k-1]$.

Require $\lambda > 0, \eta > 0$

$s = 0$, initialize $\mathbf{f}^{(s)} = \{\mathbf{f}^{l,(s)}\}_{l=1}^L$

while not converged **do**

$s = s + 1$

compute all $\nabla \mathcal{L}(\mathbf{f}^{(s)})$

determine λ by line search

for $j = 1$ to N **do**

for $l = 1$ to L **do**

$$\mathbf{f}_j^{l,(s)} = \mathbf{f}_j^{l,(s-1)} - \eta \nabla_{\mathbf{f}_j^l} \mathcal{L}(\mathbf{f}^{(s-1)})$$

end for

for $i = 0$ to $k-1$ **do**

if $\|\mathbf{f}_{(j,i)}^{1,(s)}\|_F > \lambda\eta$ **then**

$$\mathbf{f}_{(j,i)}^{1,(s)} = \mathbf{f}_{(j,i)}^{1,(s)} - \lambda\eta \frac{\mathbf{f}_{(j,i)}^{1,(s)}}{\|\mathbf{f}_{(j,i)}^{1,(s)}\|_F}$$

else

$$\mathbf{f}_{(j,i)}^{1,(s)} = 0$$

end if

end for

end while

return $\mathbf{f}^{(s)}$

the ISTA acting as the proximal operator for the Group Lasso, the group soft-thresholding operator $\mathbf{GS}_{\lambda\eta}$ for the parameters in group j is given by [44], [50]

$$\text{Prox}_{\lambda\eta}(\mathbf{w}_j) = \mathbf{GS}_{\lambda\eta}(\mathbf{w}_j)$$

$$\triangleq \begin{cases} \mathbf{w}_j - \lambda\eta \frac{\mathbf{w}_j}{\|\mathbf{w}_j\|_2}, & \|\mathbf{w}_j\|_2 > \lambda\eta \\ 0, & \|\mathbf{w}_j\|_2 \leq \lambda\eta \end{cases},$$

$$\forall \mathbf{w}_j \in \mathbb{R}^N, \quad j \in [1, N], \quad (27)$$

where λ is the hyperparameter controlling sparsity, and $\mathbf{GS}_{\lambda\eta}$ iteratively updates the weights of all N groups.

Therefore, in the channel-wise group j , the convolutional kernels in the first layer of sTCN, starting from the first kernel up to index i are denoted as $\{\mathbf{f}_{(j,i)}^1\} = [f_{(j,0)}^1, \dots, f_{(j,i)}^1]$. The hierarchical soft-thresholding operator $\mathbf{HS}_{\lambda\eta}$ for $\{\mathbf{f}_{(j,i)}^1\}$ is defined as follows:

$$\mathbf{HS}_{\lambda\eta}(\mathbf{f}_{(j,i)}^1) := \begin{cases} \mathbf{f}_{(j,i)}^1 - \lambda\eta \frac{\mathbf{f}_{(j,i)}^1}{\|\mathbf{f}_{(j,i)}^1\|_F}, & \|\mathbf{f}_{(j,i)}^1\|_F > \lambda\eta \\ 0, & \|\mathbf{f}_{(j,i)}^1\|_F \leq \lambda\eta \end{cases}, \quad \forall \mathbf{f}^1 \in \mathbb{R}^{N \times k}, \quad i \in [0, k-1] \quad (28)$$

where $\mathbf{HS}_{\lambda\eta}$ iteratively updates all k kernels in the first layer of channel j . During the i^{th} iteration, all kernels ranging from the first to the index i are updated. As a result, $\mathbf{HS}_{\lambda\eta}$ hierarchically updates all nested kernel groups as the iteration progresses. Finally, $\mathbf{HS}_{\lambda\eta}$ iteratively updates the first layer kernels $\mathbf{f}_j^1$ across all N channel-wise groups.

Therefore, the proximal gradient descent using hierarchical ISTA in the first layer, based on the backpropagation of sTCN, can be described as follows:

$$\mathbf{f}_{(j,i)}^{1,s} = \mathbf{HS}_{\lambda\eta}^{(s)} \left(\mathbf{f}_{(j,i)}^{1,(s-1)} - \eta \nabla_{\mathbf{f}_{(j,i)}^1} \mathcal{L}(\mathbf{f}^{(s-1)}) \right),$$

$$\forall \mathbf{f}^1 \in \mathbb{R}^{N \times k}. \quad (29)$$

The full optimisation algorithm is outlined in Algorithm 1. In this algorithm, the gradient of the smooth part of the objective function is first computed through backpropagation. Then, sTCN takes a smooth gradient descent step. Subsequently, the first layer of sTCN takes a proximal gradient descent step using hierarchical ISTA. This process is repeated until the convergence criterion is met.

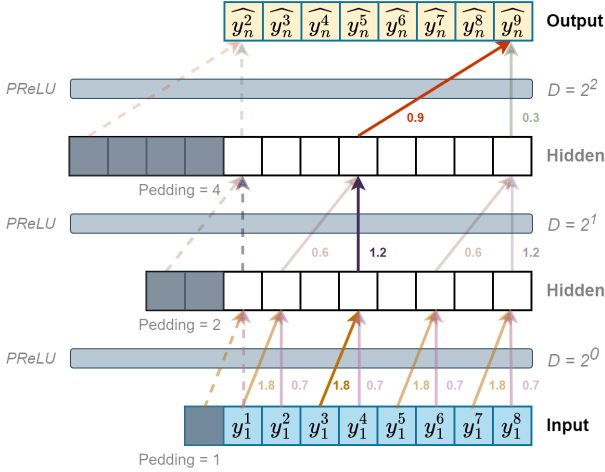


Fig. 5. An illustration depicting the process of estimating the maximum time delay between the Granger causal input y_1 and the target output y_n . The actual maximum time delay is determined by searching along the trajectory characterised by the largest kernel weights.

D. Maximum Time Delay Estimation

Following the training process, inputs with nonzero kernel weights in the first layer are identified as Granger-causal to the output time series. The maximum time delay for each causal input can be determined by searching along the trajectory characterised by the largest kernel weights, thanks to the dilation structure of sTCN [45]. The searching process is visually demonstrated in Fig. 5.

In Fig. 5, if the time series y_1 is identified as the Granger-causal input to the target output time series y_n , the delay determination algorithm starts the search for the actual maximum time delay caused by y_1 to y_n . For an sTCN consisting of 3 Conv1d layers with a kernel size of 2, each element in the output series has a receptive field of 8, according to (5). Therefore, the estimation of

the actual maximum time delay is conducted within an 8-time-step window of the input time series. The delay determination algorithm involves comparing the absolute values of all kernels in each layer to identify the largest one. It then traverses all Conv1d layers, starting from the last layer and progressing to the first, following the trajectory characterised by the largest kernel weights. To determine the actual maximum time delay, we add 1 extra step, as all inputs of the sTCN are shifted by one time step to the past. In Fig. 5, by following the trajectory characterised by the largest kernel weights and imposing an additional time delay, the actual maximum time delay that y_1 causes y_n is found to be 7.

IV. EXPERIMENTS AND RESULTS

We have conducted multiple experiments to evaluate the performance of the proposed ITCF in detecting Granger causal relationships within multivariate time series. The overall performance in terms of causality detection and time delay estimation was firstly evaluated using two simulation datasets: one representing a linear multivariate scenario and the other representing a nonlinear chaotic scenario. Then, we tested the proposed approach on more challenging datasets, including a functional magnetic resonance imaging (fMRI) dataset with limited time series length and a gene expression dataset characterized by both limited time series length and a large number of components. We compare the proposed ITCF with three baseline approaches that can also detect GC in an end-to-end manner, including two Lasso based approaches cMLP, cLSTM [42] and a TCN based approach TCDF [43]. The evaluation of the GC detection performance was measured by the micro average Area Under the Receiver Operating Characteristic curve (AUROC) which aggregates GC detection performance across all N^2 time series pairs in multivariate time series with N

TABLE I
PERFORMANCE COMPARISON OF GC DETECTION AND
TIME LAG ESTIMATION USING VAR(3) SIMULATION DATA
WITH DIFFERENT TIME SERIES LENGTH

T	Model	AUROC	Lag ACC
250	cMLP	0.91 ± 0.05	0.45 ± 0.02
	cLSTM	0.71 ± 0.04	NA
	TCDF	0.73 ± 0.03	0.78 ± 0.03
	ITCF	0.95 ± 0.04	1.00 ± 0.00
500	cMLP	0.93 ± 0.04	0.48 ± 0.02
	cLSTM	0.79 ± 0.13	NA
	TCDF	0.76 ± 0.06	0.79 ± 0.02
	ITCF	0.98 ± 0.02	1.00 ± 0.00
1000	cMLP	0.96 ± 0.02	0.55 ± 0.01
	cLSTM	0.81 ± 0.09	NA
	TCDF	0.78 ± 0.03	0.81 ± 0.02
	ITCF	1.00 ± 0.00	1.00 ± 0.00

The mean and standard deviations of AUROC and lag selection accuracy (ACC) are evaluated across 5 independent VAR(3) simulations.

components. The ROC curve demonstrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) obtained in GC detection across all N^2 pairs of time series. For cMLP and cLSTM, ROC curves were derived by sweeping across a range of regularization parameters. Similarly, the ROC curve of ITCF was obtained by sweeping λ in (8) across a range of values. The ROC curve of TCDF was derived by varying the threshold of attention scores. The AUROC generally evaluates the overall GC detection performance excluding the effect of the lambda or threshold selection. The list of tuned hyperparameters for all approaches used across different datasets is provided in the appendix. The ITCF in this study were developed with the deep learning framework PyTorch 1.11, and all data generation was conducted in Python 3.8. The GPU is NVIDIA Tesla A100 with 80GB of graphic memory. Our implementation is available on: <https://github.com/dexter-dong/ITCF>.

A. VAR simulation data experiment

In the first experiment, the multivariate time series Y intended for Granger causality analysis was generated from a 3rd order VAR model denoted as VAR(3), which is defined as follows:

$$Y^t = A_1 Y^{t-1} + A_2 Y^{t-2} + A_3 Y^{t-3} + e_t, \quad (30)$$

where Y consists of 10 components $[y_1, \dots, y_{10}]$ and the regression coefficient matrices $A_i, i = 1, 2, 3$ describe the linear interactions between the past information of these 10 components. We randomly generated the position of non-zero coefficients in each matrix with a sparsity of 20%, while making all matrices adhere to the same sparsity pattern, which created a sparse causal network within Y [54]. All coefficients were randomly generated in accordance with the eigenvalue set bounded within the unit circle in order to ensure the stationarity of Y . Additionally, e_t represents Gaussian white noise with a mean of zero and a variance of 0.01. Fig. 6 displays a sample of the initial 100 time points for two components within the VAR(3) dataset.

The impact of varying time series length on model performance in GC detection and time lag selection was analysed with the time series length T set at 250, 500, and 1000. The performance was evaluated across 5 independent simulations of VAR(3) to ensure reliability in the results. Table

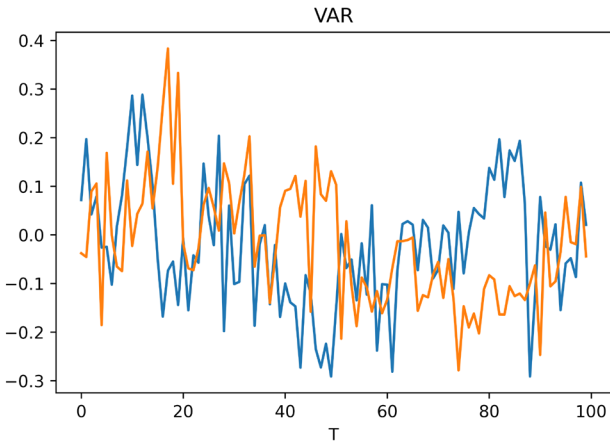


Fig. 6. The illustration of the initial 100 realizations of two distinct components within the VAR(3) dataset.

TABLE II
THE COMPARISON OF AUROC BETWEEN THE PROPOSED CHGL AND THE CONVENTIONAL GL METHOD WHEN APPLIED TO THE ITCF FOR GC DETECTION IN VAR SIMULATION DATASET WITH VARIOUS ORDERS

Lasso type	VAR order				
	5	10	13	15	18
GL	0.96 ± 0.02	0.74 ± 0.05	0.63 ± 0.03	0.59 ± 0.04	0.58 ± 0.07
chGL	1.00 ± 0.00	0.97 ± 0.03	0.81 ± 0.02	0.77 ± 0.05	0.75 ± 0.03

For each VAR order, the mean and standard deviations of AUROC were evaluated across 5 independent simulations.

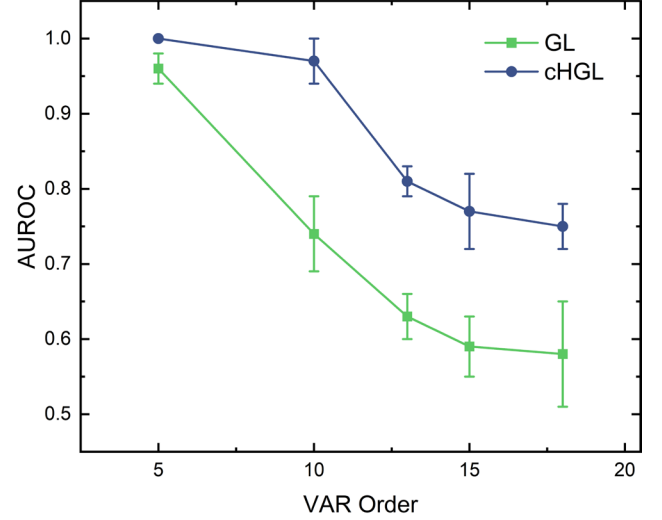


Fig. 7. A graph depicting the mean AUROC of the proposed chGL and the conventional GL method when applied to the ITCF for GC detection across various VAR Orders. Error bars represent the standard deviations.

I shows that as the available time series length increases, the GC detection performance of all models improves. cMLP outperformed cLSTM and TCDF due to its utilisation of temporal information in GC detection through the HGL, aligning with the findings in [23], [24], [44], [52]. Compared to other models, the proposed ITCF demonstrates superior

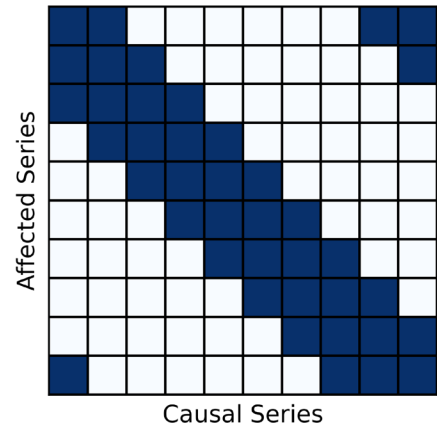


Fig. 8. The adjacency matrix representing the GC relationships in the Lorenz-96 dataset. The dark blue cells represent the existence of GC relationships.

TABLE III
THE COMPARISON OF AUROC FOR GC DETECTION AMONG DIFFERENT METHODS USING LORENZ-96 SIMULATION DATASET
WITH DIFFERENT TIME SERIES LENGTH T AND FORCING TERM F

Model	$F = 10$			$F = 40$		
	T = 250	T = 500	T = 1000	T = 250	T = 500	T = 1000
cMLP	0.83 ± 0.03	0.95 ± 0.02	0.97 ± 0.02	0.82 ± 0.05	0.88 ± 0.04	0.93 ± 0.04
cLSTM	0.76 ± 0.09	0.92 ± 0.05	0.95 ± 0.01	0.73 ± 0.08	0.85 ± 0.06	0.90 ± 0.03
TCDF	0.70 ± 0.01	0.72 ± 0.04	0.86 ± 0.03	0.65 ± 0.03	0.69 ± 0.04	0.75 ± 0.04
ITCF	0.92 ± 0.02	0.98 ± 0.02	1.00 ± 0.00	0.83 ± 0.03	0.89 ± 0.02	0.91 ± 0.02

The mean and standard deviations of AUROC were evaluated across 5 independent simulations.

performance in GC detection across all lengths of time series. Furthermore, the proposed ITCF can accurately estimate the appropriate time delay for time series of all lengths, outperforming other methods significantly. It also facilitates GC detection of the ITCF. On the other hand, because each time step of input time series shares the same set of weights in the cLSTM, it is not possible to estimate the appropriate time delay [44]. Consequently, lag selection accuracy of cLSTM is marked as 'NA' in Table I. For the cMLP, the performance deterioration in time delay detection is probably attributed to its fully connected hidden layers, which make it susceptible to overfitting specific temporal patterns. As for TCDF, the channel-wise attention score applied to the input time series might affect the magnitude of the kernel weights, potentially deteriorating delay discovery through weight interpretation.

Then, we evaluated the performance of the proposed cHGL method in comparison to the conventional GL methods when applied to TCN for detecting GC in VAR simulation datasets. We considered various VAR orders, ranging from 5 to 18. As the order of VAR increases, GC detection becomes increasingly susceptible to overfitting on certain causal inputs, posing a greater challenge. The performance was evaluated across 5 independent simulations of each VAR order, and the results are presented in Table II and Fig. 7.

In Fig. 7, the cHGL model consistently demonstrated superior performance over GL, maintaining a relatively high AUROC value even increasing VAR orders. This trend persisted across VAR orders, with cHGL outperforming GL by significant margins. Notably, at a VAR order of 18, cHGL retained a robust AUROC of around 0.75, whereas the GL model's AUROC diminished further to around 0.58. This suggests that cHGL exhibits greater resilient to the increase in the VAR order. Therefore, the proposed cHGL is demonstrated to be less prone to overfitting than the GL. This advantage primarily arises from the fact that, in contrast to the GL, the cHGL incorporates temporal information during causal input detection.

B. Lorenz-96 simulation data experiment

In the second experiment, the multivariate time series Y for Granger causality analysis was generated from the Lorenz-96 model which is a nonlinear model widely used in the climate system, as given by [55]

$$\frac{dY_i^t}{dt} = (Y_{(i+1)}^t - Y_{(i-2)}^t)Y_{(i-1)}^t - Y_i^t + F, \quad 1 \leq i \leq N, \quad (31)$$

where the multivariate time series $Y = [y_1, \dots, y_{10}]$ consists of 10 components. In our setting, the i^{th} time series in Y is caused by the time series indices from $i - 2$ to $i + 1$ and F represents the forcing term. The adjacency matrix representing the ground

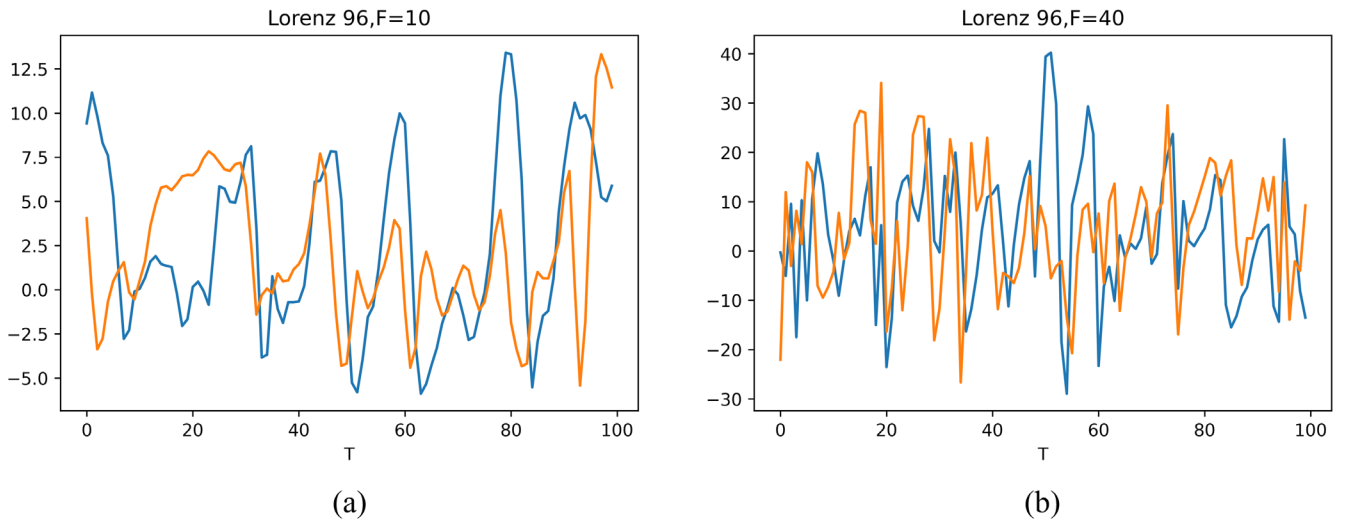


Fig. 9. The illustration of the initial 100 realisations of two distinct components in the Lorenz-96 dataset under two chaotic scenarios: (a) realisations of two distinct components under a weak chaotic scenario with $F = 10$, (b) realisations of two distinct components under a highly chaotic scenario with $F = 40$.

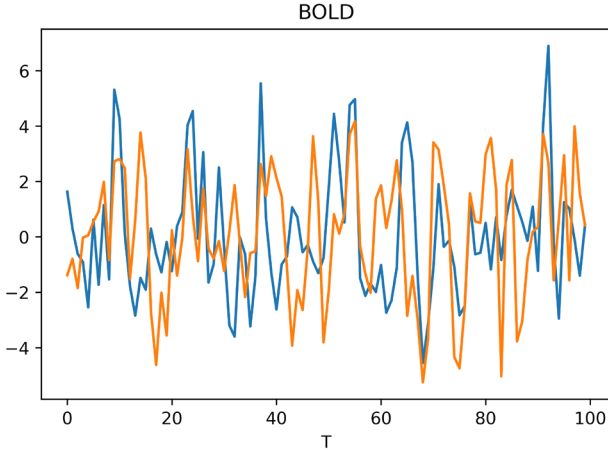


Fig. 10. The illustration of the initial 100 realizations of two distinct ROIs within the BOLD dataset.

TABLE IV
PERFORMANCE COMPARISON OF GC DETECTION BY
ESTIMATING THE BRAIN FUNCTIONAL CONNECTIVITY WITH
THE BOLD DATASET

Model	AUROC
	T = 200
cMLP	0.80 ± 0.05
cLSTM	0.71 ± 0.04
TCDF	0.77 ± 0.03
ITCF	0.83 ± 0.03

The mean and standard deviations of AUROC were evaluated across 5 human subjects.

truth GC relationships in Lorenz-96 dataset is depicted in Fig. 8.

The Lorenz-96 model starts to exhibit chaotic behaviour when $F \geq 8$ and it becomes increasing chaotic as F increases. In the experiment, we evaluated all approaches under two different scenarios: a less chaotic scenario with $F = 10$ and a more chaotic scenario with $F = 40$. Fig. 9 illustrates the initial 100 realisations of two distinct components in the Lorenz-96 simulation under these two chaotic scenarios. In Fig. 9, it has been observed that the dynamics of components become more complicated at $F=40$ compared to $F=10$. Since chaotic systems are highly sensitive to initial conditions, where tiny variations in the initial condition can lead to vastly different behaviours, we assessed all approaches in the experiment using 5 independent random initializations.

We evaluated the performance of all approaches across various time series lengths in both chaotic scenarios, presenting the results in Table III. In both chaotic scenarios, the GC detection performance of all models improves with longer available time series lengths. cMLP generally outperformed cLSTM across all lengths of time series in both chaotic scenarios, which echoes the findings in [44]. It further implies that the incorporation of temporal information into regularisation could improve the GC detection performance. In the less chaotic scenario, it has been observed that the proposed ITCF outperformed other methods across all time series lengths. However, TCDF exhibits a deteriorated performance

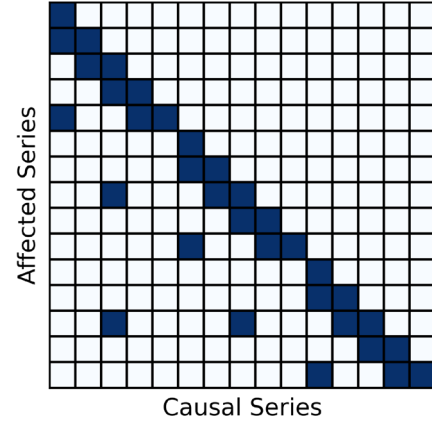


Fig. 11. The adjacency matrix representing the GC relationships in the BOLD dataset. The dark blue cells represent the existence of GC relationships.

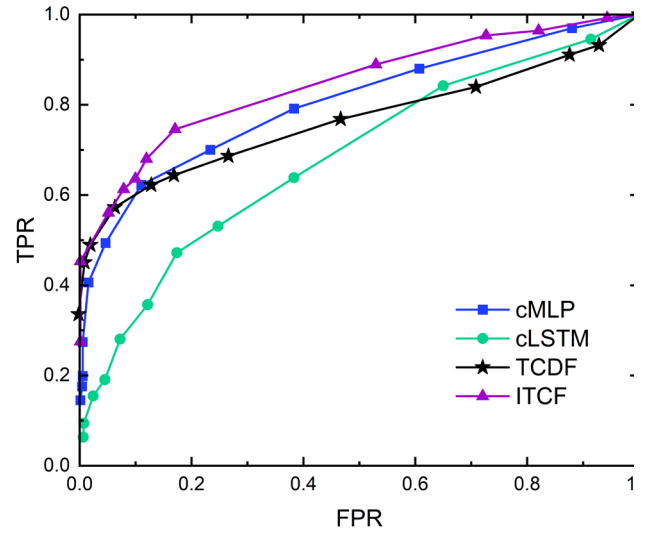


Fig. 12. The ROC curves of the brain functional connectivity estimation with the BOLD dataset. The mean TPR and FPR of each method were evaluated across 5 human subjects.

even with a long time series length of 1000. In the highly chaotic scenario, ITCF performs better than other methods at the time series lengths of 250 and 500. Surprisingly, cMLP outperformed ITCF with a longer time series length of 1000,

TABLE V
SENSITIVITY TEST OF ITCF WITH DIFFERENT NUMBER OF
HIDDEN LAYERS WITH THE BOLD DATASET

Number of hidden layers	AUROC
	T = 200
H = 1	0.83 ± 0.03
H = 2	0.83 ± 0.06
H = 3	0.81 ± 0.04
H = 4	0.80 ± 0.03

The mean and standard deviations of AUROC were evaluated across 5 human subjects.

TABLE VI
THE AUROC COMPARISON OF GC DETECTION BY
INFERRING GENE REGULATORY NETWORKS WITH 5
DISTINCT GENE EXPRESSION DATASETS IN DREAM-3
CHALLENGE

Model	AUROC				
	E.coli-1	E.coli-2	Yeast-1	Yeast-2	Yeast-3
cMLP	0.623	0.572	0.573	0.508	0.528
cLSTM	0.642	0.639	0.588	0.511	0.536
TCDF	0.602	0.603	0.585	0.535	0.547
ITCF	0.625	0.622	0.593	0.576	0.565

likely due to its capacity, with more parameters, to capture complicated behaviours in longer time series. On the other hand, the discrepancy in performance between TCDF and the proposed ITCF is more pronounced compared to the less chaotic scenario, which indicates that incorporating temporal information in the GC detection becomes increasingly beneficial for the performance in more complicated scenarios.

C. Brain functional connectivity estimation challenge

In the third experiment, we evaluated the GC detection performance of all methods by estimating the brain functional connectivity with the blood-oxygen-level-dependent (BOLD) dataset [56]. The BOLD dataset was generated from the realistic simulation of functional magnetic resonance imaging (fMRI) models, where each time series represents the activity of a specific region of interest (ROI) within the human brain. The BOLD dataset comprises 15 ROIs for each human subject, where each ROI features a time series length of 200. The relatively limited time series length poses a challenge for GC detection. Fig. 10 presents an illustration of the initial 100 realizations of two distinct ROIs within the BOLD dataset, where the dynamics are evidently complex. All human subjects share the same ground truth GC relationships among 15 brain ROIs, as shown in Fig. 11. The objective is to detect the GC among the various ROIs within the brain.

In this experiment, the mean and standard deviations of AUROC were evaluated across 5 human subjects (labelled as 0 to 4). The results are shown in Table IV. Additionally, the corresponding ROC curves for GC detection, depicting the mean TPR and FPR of each method across 5 human subjects, are illustrated in Fig. 12.

Table IV reveals that the proposed ITCF exhibits the highest mean AUROC of 0.83 with a comparatively low standard deviation of 0.03. This suggests a consistently superior performance in GC detection across the 5 human subjects. Fig. 10 reinforces the superior performance of ITCF, showcasing its ROC curve maintaining a higher TPR against FPR across all thresholds. Compared to TCDF, ITCF achieved an approximately 5% higher mean AUROC, emphasizing the benefit of incorporating temporal information into GC detection, particularly with limited time series length. On the other hand, cMLP performed only slightly less effectively than ITCF, potentially due to its simpler architecture with only one hidden layer in the hyperparameter setting. This simplicity of structure may make it less prone to overfitting even with limited time series length. In summary, the proposed ITCF has proven

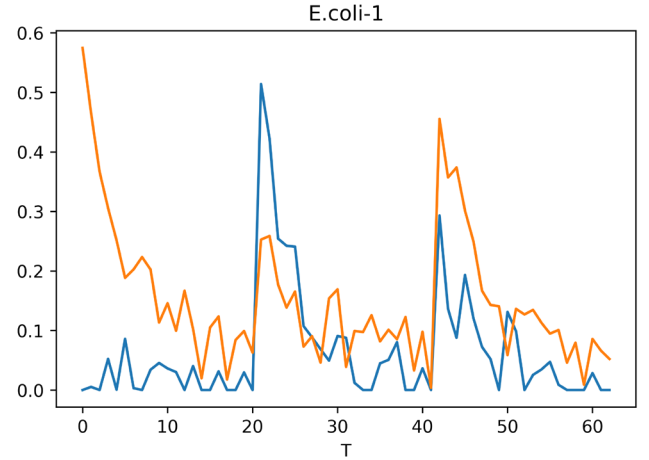


Fig. 14. The illustration of the initial 63 realizations of two distinct components within the E.coli-1 dataset.

to perform well in scenarios involving limited time series length and complex dynamics, as evidenced by both the ROC curve analysis and the AUROC scores.

Finally, the sensitivity test was carried out to evaluate the effect of the number of hidden layers, as the number of hidden layers influences the receptive field of the input time series according to (5). An excessively large receptive field may result in considering overly long time delays and may cause overfitting. The results are shown in Table V, where the number of hidden layers, ranging from 1 to 4, was tested with a maximum receptive field of 9. Although performance decreases as the number of hidden layers increases, due to the proposed cHGL and delay detection mechanism, the performance only decreases slightly by 0.3, even with a receptive field of 9. Therefore, the proposed cHGL generally performs effectively in removing redundant temporal information and preventing overfitting.

D. DREAM-3 challenge

The fourth experiment aims to evaluate the performance of GC detection by inferring gene regulatory networks from gene expression datasets characterized by numerous components and notably short time series lengths. The gene expression dataset is synthesized for the Dialogue on Reverse Engineering Assessment and Methods 3 (DREAM-3) challenge [57]. The challenge comprises five distinct simulated datasets, each including distinct ground truth GC graphs as shown in Fig. 13. The gene expression datasets consist of two E.coli (E.C.) datasets and three Yeast (Y.) datasets. Each dataset consists of 100 distinct gene components, and the trajectory of each component is recorded from their recovery from 46 different perturbations. The recovery trajectory of each perturbation has 21 time points, resulting in a cumulative total of 966 time points. Thus, the time series length is notably limited relative to the component dimension and the intricate underlying behaviours. Fig. 14 shows an illustration of the initial 63 realizations, which correspond to three perturbation cycles of two distinct components in the E.coli-1 dataset.

In this experiment, all methods were evaluated using five distinct gene expression datasets, and the results are presented in Table VI. Analysis of Table VI reveals that cLSTM

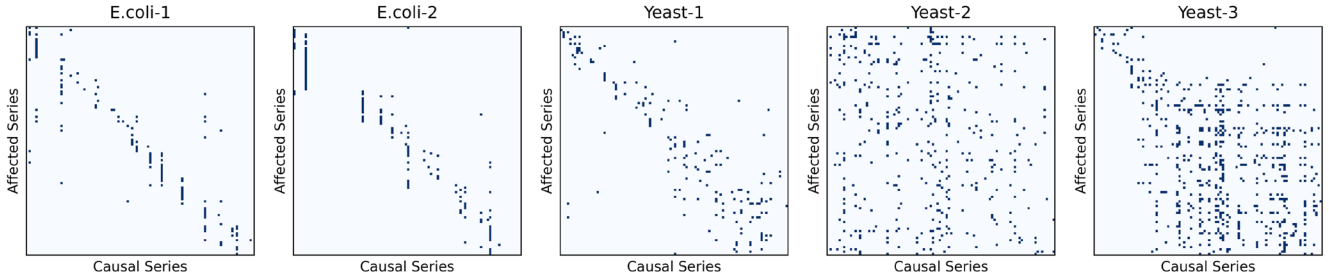


Fig. 13. The adjacency matrix representing the GC relationships in the DREAM-3 challenge. The dark blue cells represent the existence of GC relationships.

outperformed other methods in two E.coli datasets, while ITCF's performance was generally on par with it. This is probably because the cLSTM tends to select sparser causal inputs, as all weights, including the gate weights, are regularized by group Lasso. This could result in the cLSTM giving less consideration to causal inputs with longer time delays, making it suitable for simpler scenarios, such as the E.coli datasets. However, when tackling more complex scenarios with more causal inputs, such as the Yeast datasets, its performance will deteriorate. In contrast, although the ITCF is less susceptible to overfitting due to the proposed cHGL, its performance may still be affected by noise, which becomes more evident in scenarios with sparser causal relationships. On the other hand, ITCF outperformed all other methods in the three Yeast datasets with a small margin. Compared to TCDF, the proposed ITCF consistently demonstrated superior performance across all five datasets. This highlights the advantages of incorporating temporal information in GC detection, particularly in scenarios involving a large number of candidate causal inputs. In summary, the proposed ITCF maintains robust and promising performance in challenging scenarios characterized by a large number of candidate causal inputs and short time series lengths.

V. DISCUSSION

In real-world applications, the selection of lambda significantly impacts GC detection results and is crucial for Lasso-based approaches [52]. Lambda selection is typically performed using a line search over a range of lambda values, which often involves cross-validation [44]. During cross-validation, the data is split into training and validation sets, and the model is trained multiple times with different lambda values. The performance is then evaluated on the validation set to select the lambda that minimizes an error metric, such as Mean Squared Error (MSE) or Akaike Information Criterion (AIC). However, such methods are computationally intensive since the neural network must be trained multiple times for each lambda value. With the development of distributed training [58], these search steps could be parallelized across multiple GPUs to reduce computation time. Our future work will focus on developing a systematic approach for efficient lambda selection.

On the other hand, the theoretical convergence rate of ISTA is $O(1/k)$, where k is the iteration number. This acceleration reduces the number of iterations required for a given level of accuracy. By leveraging the structure of convex problems, FISTA efficiently handles smooth parts of the optimization while applying shrinkage for sparsity [51]. Furthermore, the

Alternating direction method of multipliers (ADMM), is particularly designed for problems with separable structures. It handles constraints and separable objectives by alternating between variables and using dual updates. This decomposition can lead to faster convergence in practice, especially for structured problems like LASSO [59]. However, the integration of FISTA and ADMM into the backpropagation of neural networks remains unexplored. Our future research will focus on improving the convergence rate of ISTA in backpropagation by adjusting FISTA and ADMM.

It should be noted that this paper focuses specifically on Granger causality, which is inherently based on time-series data and aims to detect temporal dependencies within the given data. We fully acknowledge that real-world systems are influenced by a variety of external factors, and relying solely on time-series data may oversimplify causal relationships. Future works will extend the framework by including external variables or other data modalities.

VI. CONCLUSION

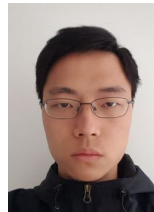
In this work, we proposed a novel Interpretable Temporal Convolutional Framework (ITCF), a temporal convolution networks (TCN)-based approach that employs sparse regression to detect both linear and nonlinear Granger Causality (GC) and automatically estimate time delay in multivariate time series. Within the ITCF, we introduced a convolutional Hierarchical Group Lasso (cHGL), a hierarchical regularisation approach facilitating GC detection by effectively utilising temporal information within each TCN channel. Additionally, we optimized the cHGL by incorporating the Iterative Soft-Thresholding Algorithm (ISTA) into the backpropagation, leading to an end-to-end framework for GC detection. The proposed ITCF generally outperformed three baselines across all time series lengths in two simulation datasets and exhibited a resilient performance in two realistic benchmarks. The results empirically demonstrate that the proposed ITCF performed promisingly in both linear and nonlinear scenarios without the need for manual time delay selection. Furthermore, the proposed cHGL enhances the ITCF's robustness and outperforms other baseline methods, especially in challenging scenarios characterised by complicated dynamics, limited time series lengths, and numerous candidate components. The promising performance indicates the potential of the proposed ITCF for more reliable analysis of GC relationships in complex systems, which could be beneficial for research in climate forecasting, neuroscience and clinical diagnosis. As the selection of lambda can affect the results of GC detection [52],

the future work will focus on developing a systematic approach for lambda selection. One limitation of the proposed ITCF is that the GC detection relies on the assumption of a time invariant nature of the underlying system, whereas time-varying systems widely exist in real-world scenarios. Future work will aim to extend the method to handle these time-varying scenarios, enhancing the method's applicability in predictive modelling.

REFERENCES

- [1] Y. Zhao, E. Hanna, G. R. Bigg, and Y. Zhao, "Tracking Nonlinear Correlation for Complex Dynamic Systems Using a Windowed Error Reduction Ratio Method," *Complexity*, vol. 2017, 2017, doi: 10.1155/2017/8570720.
- [2] A. Papana, C. Kyrtou, D. Kugiumtzis, and C. Diks, "Financial networks based on Granger causality: A case study," *Physica A*, vol. 482, pp. 65–73, 2017, doi: 10.1016/j.physa.2017.04.046.
- [3] X. R. Hou, K. Wang, C. Zhong, and Z. Wei, "ST-Trader: A Spatial-Temporal Deep Neural Network for Modeling Stock Market Movement," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 5, pp. 1015–1024, May. 2021, doi: 10.1109/JAS.2021.1003976.
- [4] J. Cao, Y. Zhao, X. Shan, H. Wei, Y. Guo, L. Chen, J. A. Erkoyuncu and P. G. Sarrigiannis, "Brain functional and effective connectivity based on electroencephalography recordings: A review," *Hum Brain Mapp*, 2021, doi: 10.1002/HBM.25683.
- [5] X. Chen and Y. Wang, "Predicting Resting-state Functional Connectivity With Efficient Structural Connectivity," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 6, pp. 1079–1088, Nov. 2018, doi: 10.1109/JAS.2017.7510880.
- [6] D. A. Smirnov and I. I. Mokhov, "From Granger causality to long-term causality: Application to climatic data," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 80, no. 1, p. 016208, Aug. 2009, doi: 10.1103/PHYSREVE.80.016208/FIGURES/13/MEDIUM.
- [7] Y. Zhao, G. R. Bigg, S. A. Billings, E. Hanna, A. J. Sole, H. Wei, V. Kadiramanathan and D. J. Wilton, "Inferring the variation of climatic and glaciological contributions to West Greenland iceberg discharge in the twentieth century," *Cold Reg Sci Technol*, vol. 121, pp. 167–178, 2016, doi: 10.1016/j.coldregions.2015.08.006.
- [8] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, p. 424, Aug. 1969, doi: 10.2307/1912791.
- [9] A. Shojaie and E. B. Fox, "Granger Causality: A Review and Recent Advances," *Annu Rev Stat Appl*, vol. 9, no. 1, pp. 289–319, Mar. 2022, doi: 10.1146/annurev-statistics-040120-010930.
- [10] T. Schreiber, "Measuring Information Transfer," *Phys Rev Lett*, vol. 85, no. 2, pp. 461–464, Jan. 2000, doi: 10.1103/PhysRevLett.85.461.
- [11] L. Faes, G. Nollo, and A. Porta, "Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique," *Phys Rev E*, vol. 83, no. 5, p. 051112, May 2011, doi: 10.1103/PhysRevE.83.051112.
- [12] M. Dhamala, G. Rangarajan, and M. Ding, "Estimating Granger Causality from Fourier and Wavelet Transforms of Time Series Data," *Phys Rev Lett*, vol. 100, no. 1, p. 018701, Jan. 2008, doi: 10.1103/PhysRevLett.100.018701.
- [13] E. Torun, T. P. Chang, and R. Y. Chou, "Causal relationship between spot and futures prices with multiple time horizons: A nonparametric wavelet Granger causality test," *Res Int Bus Finance*, vol. 52, p. 101115, Apr. 2020, doi: 10.1016/J.RIBAF.2019.101115.
- [14] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality," *Phys Rev Lett*, vol. 100, no. 14, p. 144103, Apr. 2008, doi: 10.1103/PHYSREVLETT.100.144103/FIGURES/4/MEDIUM.
- [15] S. Seth and J. C. Principe, "Assessing granger non-causality using nonparametric measure of conditional independence," *IEEE Trans Neural Netw Learn Syst*, vol. 23, no. 1, pp. 47–59, 2012, doi: 10.1109/TNNLS.2011.2178327.
- [16] Y. Zhao, S. A. Billings, H. Wei, and P. G. Sarrigiannis, "Tracking time-varying causality and directionality of information flow using an error reduction ratio test with applications to electroencephalography data," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 86, no. 5, Nov. 2012, doi: 10.1103/PhysRevE.86.051919.
- [17] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, "Escaping the curse of dimensionality in estimating multivariate transfer entropy," *Phys Rev Lett*, vol. 108, no. 25, p. 258701, Jun. 2012, doi: 10.1103/PHYSREVLETT.108.258701/FIGURES/4/MEDIUM.
- [18] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical granger modeling methods for temporal causal modeling," *Proc of the ACM Intern Conf on Know Discov and Data Mining (SIGKDD)*, pp. 577–585, 2009, doi: 10.1145/1557019.1557085.
- [19] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistical Learning with Sparsity". *Chapman and Hall/CRC*, 2015. doi: 10.1201/b18401.
- [20] M. Yuan and Y. Lin, "Model Selection and Estimation in Regression with Grouped Variables," *J R Stat Soc Series B Stat Methodol*, vol. 68, no. 1, pp. 49–67, Feb. 2006, doi: 10.1111/j.1467-9868.2005.00532.x.
- [21] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical Granger modeling for gene expression regulatory networks discovery," *Bioinformatics*, vol. 25, no. 12, 2009, doi: 10.1093/BIOINFORMATICS/BTP199.
- [22] W. B. Nicholson, I. Wilms, J. Bien, and D. S. Matteson, "High Dimensional Forecasting via Interpretable Vector Autoregression," *J. of Mach Learn Res*, vol. 21, Dec. 2014, Accessed: Oct. 16, 2023. [Online]. Available: <http://arxiv.org/abs/1412.5250>
- [23] Y. Antonacci, L. Astolfi, G. Nollo, and L. Faes, "Information Transfer in Linear Multivariate Processes Assessed through Penalized Regression Techniques: Validation and Application to Physiological Networks," *Entropy*, vol. 22, no. 7, p. 732, Jul. 2020, doi: 10.3390/e22070732.
- [24] Y. Antonacci, J. Toppi, A. Pietrabissa, A. Anzolin, and L. Astolfi, "Measuring Connectivity in Linear Multivariate Processes With Penalized Regression Techniques," *IEEE Access*, vol. 12, pp. 30638–30652, 2024, doi: 10.1109/ACCESS.2024.3368637.
- [25] S. A. Billings, *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. Chichester, UK: Wiley, 2013. doi: 10.1002/9781118535561.
- [26] Y. Zhao, S. A. Billings, H. Wei, F. He, and P. G. Sarrigiannis, "A new NARX-based Granger linear and nonlinear causal influence detection method with applications to EEG data," *J Neurosci Methods*, vol. 212, no. 1, pp. 79–86, Jan. 2013, doi: 10.1016/j.jneumeth.2012.09.019.
- [27] Y. Li, H.-L. Wei, S. A. Billings, and X.-F. Liao, "Time-varying linear and nonlinear parametric model for Granger causality analysis," *Phys Rev E*, vol. 85, no. 4, p. 041906, Apr. 2012, doi: 10.1103/PhysRevE.85.041906.
- [28] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Nonlinear parametric model for Granger causality of time series," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 73, no. 6, p. 066216, Jun. 2006, doi: 10.1103/PHYSREVE.73.066216/FIGURES/9/MEDIUM.
- [29] S. Chen, X. Hong, B. L. Luk, and C. J. Harris, "Construction of tunable radial basis function networks using orthogonal forward selection," *IEEE Trans on Sys, Man, and Cyber, Part B: Cyber*, vol. 39, no. 2, pp. 457–466, 2009, doi: 10.1109/TSMCB.2008.2006688.
- [30] W. Ren, B. Li, and M. Han, "A novel Granger causality method based on HSIC-Lasso for revealing nonlinear relationship between multivariate time series," *Physica A: Statis Mech and its App*, vol. 541, p. 123245, Mar. 2020, doi: 10.1016/j.physa.2019.123245.
- [31] C. C. Aggarwal, *Neural Networks and Deep Learning*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-94463-0.
- [32] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989, doi: 10.1016/0893-6080(89)90020-8.
- [33] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philo Trans of the Royal Soc A: Math, Phys and Eng Sci*, vol. 379, no. 2194, p. 20200209, Apr. 2021, doi: 10.1098/rsta.2020.0209.
- [34] H. Wang and G. Song, "Innovative NARX recurrent neural network model for ultra-thin shape memory alloy wire," *Neurocomputing*, vol. 134, pp. 289–295, Jun. 2014, doi: 10.1016/j.neucom.2013.09.050.
- [35] J. L. Elman, "Finding Structure in Time," *Cogn Sci*, vol. 14, no. 2, pp. 179–211, Mar. 1990, doi: 10.1207/s15516709cog1402_1.
- [36] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.
- [37] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *EMNLP 2014 - 2014 Conf on Emp Methods in Natural Lang Proc, Proc of the Conf*, pp. 1724–1734, Jun. 2014, doi: 10.3115/v1/d14-1179.
- [38] Y. Wang, K. Lin, Y. Qi, Q. Lian, S. Feng, Z. Wu and G. Pan, "Estimating brain connectivity with varying-length time lags using a recurrent neural network," *IEEE Trans Biomed Eng*, vol. 65, no. 9, pp. 1953–1963, Sep. 2018, doi: 10.1109/TBME.2018.2842769.

- [39] P. J. Werbos, "Backpropagation Through Time: What It Does and How to Do It," *Proc of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990, doi: 10.1109/5.58337.
- [40] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv preprint, arXiv:1803.01271, Mar. 2018.
- [41] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," arXiv preprint, arXiv:1609.03499, Sep. 2016.
- [42] A. Montalto, S. Stramaglia, L. Faes, G. Tessitore, R. Prevete, and D. Marinazzo, "Neural networks with non-uniform embedding and explicit validation phase to assess Granger causality," *Neur Net*, vol. 71, pp. 159–171, Nov. 2015, doi: 10.1016/J.NEUNET.2015.08.003.
- [43] B. Liu, X. He, M. Song, J. Li, G. Qu, J. Lang and R. Gu, "A Method for Mining Granger Causality Relationship on Atmospheric Visibility," *ACM Trans on Know Disc from Data (TKDD)*, vol. 15, no. 5, May 2021, doi: 10.1145/3447681.
- [44] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. B. Fox, "Neural Granger Causality," *IEEE Trans Pattern Anal Mach Intell*, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3065601.
- [45] M. Nauta, D. Bucur, and C. Seifert, "Causal Discovery with Attention-Based Convolutional Neural Networks," *Mach Learn and Know Extra*, Vol. 1, Pages 312–340, vol. 1, no. 1, pp. 312–340, Jan. 2019, doi: 10.3390/MAKE1010019.
- [46] T. Shi, W. Yang, A. Qi, P. Li, and J. Qiao, "LASSO and attention-TCN: a concurrent method for indoor particulate matter prediction," *Appl Intell*, vol. 53, no. 17, pp. 20076–20090, Sep. 2023, doi: 10.1007/S10489-023-04507-6/TABLES/5.
- [47] Y. Shao, J. Tang, J. Liu, L. Han, and S. Dong, "Multivariable System Prediction Based on TCN-LSTM Networks with Self-Attention Mechanism and LASSO Variable Selection," *ACS Omega*, Dec. 2023, doi: 10.1021/ACSOMEGA.3C06263.
- [48] M. Rosoł, M. Młyńczak, and G. Cybulski, "Granger causality test with nonlinear neural-network-based methods: Python package and simulation study," *Comput Methods Programs Biomed*, p. 106669, Jan. 2022, doi: 10.1016/J.CMPB.2022.106669.
- [49] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal Methods for Hierarchical Sparse Coding," *J. of Mach Learn Res*, vol. 12, pp. 2297–2334, Sep. 2010, Accessed: Oct. 11, 2023. [Online]. Available: <http://arxiv.org/abs/1009.2139>
- [50] N. Parikh, "Proximal Algorithms," *Found and Trends® in Optim*, vol. 1, no. 3, pp. 127–239, Jan. 2014, doi: 10.1561/24000000003.
- [51] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J Imaging Sci*, vol. 2, no. 1, pp. 183–202, Jan. 2009, doi: 10.1137/080716542.
- [52] Y. Antonacci, L. Minati, L. Faes, R. Pernice, G. Nollo, J. Toppi, A. Pietrabissa and L. Astolfi, "Estimation of Granger causality through Artificial Neural Networks: applications to physiological systems and chaotic electronic oscillators," *PeerJ Comput Sci*, vol. 7, p. e429, May 2021, doi: 10.7717/peerj-cs.429.
- [53] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal Methods for Hierarchical Sparse Coding," *J. of Mach Learn Res*, vol. 12, pp. 2297–2334, Sep. 2010, Accessed: Oct. 25, 2023. [Online]. Available: <http://arxiv.org/abs/1009.2139>
- [54] A. Bolstad, B. D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans on Sign Proc*, vol. 59, no. 6, pp. 2628–2641, Jun. 2011, doi: 10.1109/TSP.2011.2129515.
- [55] E. N. Lorenz, "Predictability – a problem partly solved," *Predic of Weather and Climate*, vol. 9780521848824, Cambridge University Press, 2006, pp. 40–58. doi: 10.1017/CBO9780511617652.004.
- [56] S. M. Smith et al., "Network modelling methods for FMRI," *Neuroimage*, vol. 54, no. 2, pp. 875–891, Jan. 2011, doi: 10.1016/j.neuroimage.2010.08.063.
- [57] R. J. Prill, D. Marbach, J. S Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. A. Bonnet and G. Stolovitzky, "Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges," *PLoS One*, vol. 5, no. 2, p. e9202, Feb. 2010, doi: 10.1371/journal.pone.00092.
- [58] M. Langer, Z. He, W. Rahayu, and Y. Xue, "Distributed Training of Deep Learning Models: A Taxonomic Perspective," *IEEE Trans on Para and Distri Sys*, vol. 31, no. 12, pp. 2802–2818, Dec. 2020, doi: 10.1109/TPDS.2020.3003307.
- [59] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein. "Distributed optimization and statistical learning via the alternating direction method of multipliers". *Found and Trends® in Mach learn*. 2011 Jul 25;3(1):1-22.



Aoxiang Dong received an MSc degree in Automation and Control from Newcastle University, Newcastle upon Tyne, UK, in 2019. He received his PhD degree in Data Science at Cranfield University, Cranfield, UK, in 2024.

His research interests include time series analysis, nonlinear Granger Causality analysis, and explainable artificial intelligence.



Andrew Starr received a PhD degree in condition-based maintenance from the University of Manchester, Manchester, UK, in 1993. He is currently a Professor of Maintenance Systems at Cranfield University, Cranfield, UK.

His research interests include machine and structural damage detection, diagnostics and prognostics, autonomous inspection, "Big Data" for system monitoring and control, and applications in railway infrastructure and vehicles.



Yifan Zhao received a PhD degree in Automatic Control and System Engineering from the University of Sheffield, Sheffield, UK, in 2007. He is currently a Professor of Data Science at Cranfield University, Cranfield, UK. He is a Senior Member of IEEE and a Fellow of Higher Education Academy.

His research interests include machine learning, computer vision, signal processing, pattern recognition, and nonlinear system identification.



DSPACE

<https://dspace.org/>

An interpretable temporal convolutional framework for Granger causality analysis

Dong, Aoxiang; Starr, Andrew; Zhao, Yifan

2026-03

Institute of Electrical and Electronics Engineers (IEEE)

<https://doi.org/10.1109/jas.2025.125396>

Dong A, Starr A, Zhao Y. (2026) An interpretable temporal convolutional framework for Granger causality analysis. IEEE/CAA Journal of Automatica Sinica, Volume 13, Issue 3, March 2026, pp. 665-679