SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2024, Vol. 67, Iss. 5, 152103:1–152103:13
https://doi.org/10.1007/s11432-021-3724-0

Granger causal representation learning for groups of time series

Ruichu CAI[1,2], Yunjin WU[1], Xiaokai HUANG[1], Wei CHEN[1*], Tom Z. J. FU[1] & Zhifeng HAO[1,3]

[1]*School of Computer Science, Guangdong University of Technology, Guangzhou* 510006*, China;*
[2]*Peng Cheng Laboratory, Shenzhen* 518066*, China;*
[3]*College of Science, Shantou University, Shantou* 515063*, China*

Received 12 November 2021/Revised 21 October 2022/Accepted 7 March 2023/Published online 1 April 2024

Abstract Discovering causality from multivariate time series is an important but challenging problem. Most existing methods focus on estimating the Granger causal structures among multivariate time series, while ignoring the prior knowledge of these time series, e.g., the group of the time series. Focusing on discovering the Granger causal structures among groups of time series, we propose a Granger causal representation learning method to solve this problem. First, we use the multiset canonical correlation analysis method to learn the Granger causal representation of each group of time series. Then, we model the Granger causal relationships among the learned Granger causal representations using a recurrent neural network with temporal information. Finally, we formulate the above two stages into one unified optimization problem, which is efficiently solved using the augmented Lagrangian method. We conduct extensive experiments on synthetic and real-world datasets to validate the correctness and effectiveness of the proposed method.

Keywords Granger causal discovery, Granger causal representation learning, time series data, recurrent neural network, multiset canonical correlation analysis

1 Introduction

Discovering causality from multivariate time series has attracted much attention in many fields, such as social network analysis [1], biology [2], and neuroscience [3]. Such time series data contain multiple time series over measured variables. In some cases, these time series in the low level are usually clustered into high-level groups according to certain rules, and researchers are interested in recovering causal relationships among groups of time series. For example, in computer vision, researchers care more about the causal relationship among the objects (high-level variables) in an image than that among pixels (low-level variables). In neuroscience, researchers are more interested in recovering the Granger causal graphs among regions of interest (ROIs) than that among voxels.

The typical methods for discovering the causal structure from time series are the non-Gaussian structural vector autoregressive model [4], time series models with independent noise (TiMINo) algorithm [5], Granger causality [6] and its nonlinear extensions [7, 8]. These methods focus only on the causal relationships between time series (i.e., measured variables). If these time series are collected from multiple groups, and the time series in the same group are highly correlated and high-dimensional, then researchers are more interested in the causal relationships among the representations of the group. In such cases, applying existing methods to that type of data will probably lead to high computational complexity and result in a dense causal graph. For example, in neuroscience, the number of BOLD signal data of voxels (measured time series) obtained using functional magnetic resonance imaging (fMRI) is more than thousands, and the relationships between different voxels are too dense to understand how the brain works. Instead, researchers are more interested in grouping multiple voxels into ROIs as representations and recovering the Granger causal graphs among ROIs. Based on the above discussions, if we only learn the dense causal graphs among measured variables, then not only is the meaning of these many measured

* Corresponding author (email: chenweiDelight@gmail.com)

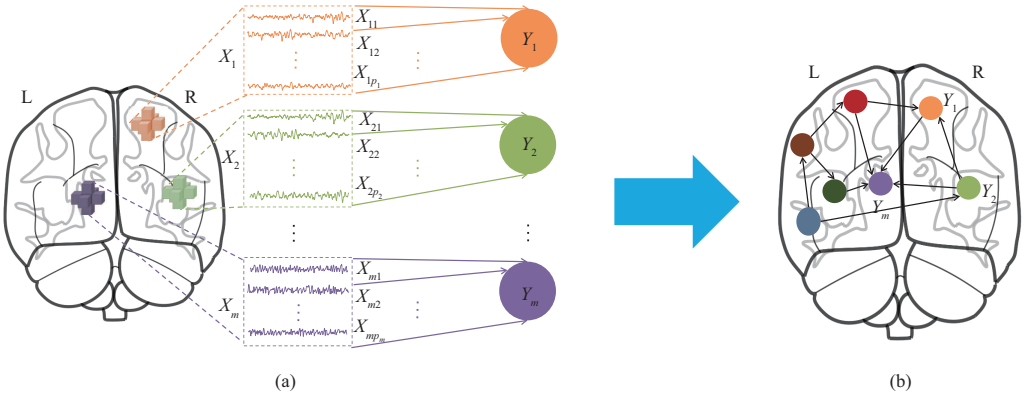

Figure 1 (Color online) Example of learning the Granger causal graph among groups of time series. (a) Multi-group time series from fMRI; (b) causal graph among variable groups.

variables and the learned relationships between the variables from the same group difficult to explain but also the rich hierarchical information between the representations and the measured variables is ignored. Thus, in this study, we aim to estimate the Granger causal graph among the representations from multigroup time-series data, such as Figure 1(b).

However, causal discovery among groups of time series is nontrivial task because of the following two challenges: (1) how to properly represent the rich temporal information from the measured variables and the variable group mapping information; and (2) how to accurately estimate the causal structure among representations from information carried by these measured variables. Aiming at these challenges, two types of methods have been proposed to learn the Granger causal structure among representations from low-level measured variables. One type of method takes the mean value of measured variables belonging to the same group as representations and then estimates the causal structure [9]. The rich temporal information in the raw time series data can be largely underused by taking an average, hence leading to inaccurate causal structure estimation [10]. The other type of method, in a different manner, first estimates the causal structure among all the measured variables regardless of which group they belong to, and further refines the causal structure according to the variable group mapping relationship [11]. This type of method has a potential defect because the estimated causal structure among all measured variables in the first step can be very dense. Even after refinement in the second step, the resulting causal structure may still be a fully connected graph, which is unwanted.

In this study, we aim to tackle these problems by proposing a Granger causal representation learning method based on the hierarchical Granger causal representation model (HGCRM) to discover the Granger causal structure among representations from multigroup time series. In detail, we learn the Granger causal representation of each group of time series using multiset canonical correlation analysis (MCCA). Meanwhile, we estimate the Granger causal relationships among representations using a recurrent neural network with temporal information. Subsequently, these two steps are nested into an optimization objective function and solved simultaneously using the augmented Lagrangian method.

The contributions of this paper are summarized below.

• We propose a hierarchical Granger causal model to learn Granger causal structures among groups of variables from multigroup time series.

• To formalize a Granger causal model for multiple variable groups, we introduce the MCCA with the causal relationship weight to learn the Granger causal representations from the groups of variables.

• We provide a unified objective function as well as an effective optimization method to learn the causal representation and the causal structures simultaneously.

• We conduct experiments to validate the correctness and effectiveness of our method on synthetic and real-world datasets.

2 Related work

Three kinds of methods are used to discover causality from time series data. The first type of method is constraint-based. Many researchers tried to generalize the Peter-Clark (PC) algorithm [12] to time

series data. Zhang et al. [13] extended the constraint-based causal discovery method to time series data. Chu et al. [14] proposed an independence-based procedure for learning nonlinear time series structures. Runge et al. [15] proposed the PCMCI (PC with momentary conditional independence) method, which comprises two stages similar to PC and uses the momentary conditional independence test to determine whether causal independence holds, which can be used in linear or nonlinear cases. However, this type of method is still affected by the Markov equivalent class problem. The second type of method is Granger causality-based. Granger causality [6] is used to predict whether time series X is the cause of another time series Y based on the autoregressive model, which can only be used in linear systems. Recently, many existing studies [16,17] have extended the Granger causality to the multivariable case in a nonlinear system. Ashrafulla et al. [16] proposed a canonical Granger causality to determine the causal relationship between two measured variable sets for the corresponding ROIs, but it only applies to the case of two sets of variables. Leveraging the acyclicity constraint, the DYNOTEARS method [17] was proposed to solve the structure learning from time series data, which can handle the case of high-dimensional variables. The last type of method is the noise independence-based, which leverages the independence between noise and causes to identify causes and effects among the given measured variables. Sanchez-Romero et al. [3] used the non-Gaussian property of the BOLD signal and provided two approaches, fast adjacency skewness (FASK) and two-step, to recover the cyclic causal network of ROIs from the preprocessed data. However, the preprocessing phase easily leads to distortions in the distributions, thus eliminating some or sometimes all of the non-Gaussianity of variables [18]. For nonlinear systems on time series, the TiMINo algorithm [5] was proposed as a typical additive noise model and can be used to discover the instantaneous effects.

The existing methods were designed for only addressing multiple time series without considering the grouping information and learning the representations. Hence, they poorly reveal the true Granger causal structure among representations. Moreover, simply clustering variables in different groups cannot use the dependence between some variable pairs belonging to the same group, which leads to poor performance on accurate Granger causal discovery. Motivated by these flaws, in this paper, we propose the HGCRM, a new methodology for tackling these challenges. It is designed to handle multiple time series and fully use variable grouping information as well as the rich hierarchical information carried by the raw input time series data.

3 Problem formulation

Let $\boldsymbol{X}$ denote the multigroup time series, and $\boldsymbol{X}_i$ denote the i-th variable group containing p_i measured variables, i.e., $\{X_{i,1}, X_{i,2}, \ldots, X_{i,p_i}\}$. Suppose that the number of variable groups is m, and the length of each time series is T. The representation of the i-th variable group is denoted as $Y_i \in \mathbb{R}^{1\times T}$. Note that Y_i can be a multi-dimensional variable. But in this paper, we consider that from the low-level variables to the high-level representation, the dimension is reduced. This assumption is also consistent with many real-world applications. For example, in neuroscience, an ROI contains many voxels; in computer vision, an object in the image consists of many pixels.

Based on the above notations, we assume that there is no direct edge between $X_{i,1}$ and $X_{j,1}$, and $X_{i,1}$ and $X_{j,1}$ are independent conditional on Y_i and Y_j. We focus on the Granger causal relationships between groups without instantaneous effect, which is defined as follows.

Definition 1. Suppose there are multiple groups of time series, each time series $\{X_{i,1}, X_{i,2}, \ldots, X_{i,p_i}\}$ within a group can be represented as a high-level causal variable Y_i, then the Granger causality between groups is that between high-level causal variables Y_i and Y_j.

According to the definition, all Granger causal relationships can be represented as a directed Granger causal graph as shown in Figure 1(b). In the Granger causal graph, each node is a Granger causal variable that represents a variable group and each edge indicates the Granger causal relationship between two causal variables. The goal is to discover the Granger causal relationships among all the m groups from $\boldsymbol{X}$. The problem can be formalized as Definition 2.

Definition 2. Problem definition: given m groups of time series data and group labels of each time series, where the i-th variable group containing p_i time series, we aim to learn the representations for groups of time series and further discover the Granger causal structure among the representations.

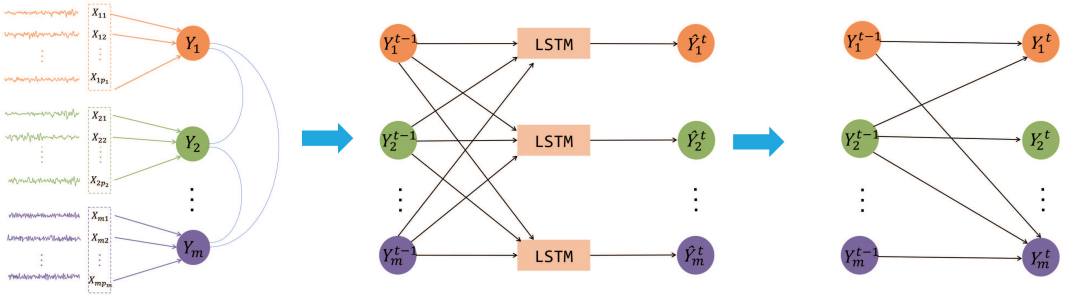

Figure 2 (Color online) Overview of the proposed model. (a) Representation of variable groups; (b) architecture of solution; (c) learned causal graph.

4 HGCRM

In this section, we will introduce an HGCRM. First, we propose a Granger causal representation learning method for measured multigroup time series. Then, we devise a Granger causal discovery on the representations. Finally, we summarize the above two steps into a unified model.

4.1 Granger causal representation learning on variable groups

In this subsection, we dive into the learning of the representations from multigroup time series. Inspired by [19], we try to connect the latent representation Y_i to the corresponding group of variables $\{X_{i,1}, X_{i,2}, \ldots, X_{i,p_i}\}$ as

$$Y_i = F(X_{i,1}, X_{i,2}, \ldots, X_{i,p_i}) \quad (i = 1, 2, \ldots, m), \tag{1}$$

where F is a mapping function. An example is given in Figure 2(a), where the high-level representation is extracted from the low-level observations for each group.

The traditional methods [3,9,20] are to use the mean of the variables in the group as the representation of this group, i.e., if the representation of the i-th group containing p_i measured variables at the time t is denoted by $Y_i^{(t)}$, then $Y_i^{(t)}$ can be obtained by the following equation:

$$Y_i^{(t)} = \frac{1}{p_i} \sum_{k=1}^{p_i} X_{i,k}^{(t)}, \tag{2}$$

where $X_{i,k}^{(t)}$ is the k-th measured variables in i-th group at the time t. However, it can be shown that the rich hierarchical information may be lost by averaging [10], which also makes subsequent steps in estimating the Granger causal graph infeasible.

In contrast to the above simple method, we investigate another Granger causal representation method to represent multigroup variables. This is related to the idea of the MCCA [21]. The MCCA represents each group with the corresponding canonical variable $Y_i^{(t)}$ at the time t defined as follows:

$$Y_i^{(t)} = A_i^{*\mathrm{T}} X_i^{(t)}, \tag{3}$$

where $X_i^{(t)}$ are the measured variables in the i-th group at the time t, and A_i^* is the optimal vector that is obtained by solving the following optimization problem:

$$\begin{aligned} &A_1^*, A_2^*, \ldots, A_m^* = \arg\max \sum_{i=1}^{m} \sum_{j=i+1}^{m} A_i^{\mathrm{T}} \Sigma_{\boldsymbol{X}_i \boldsymbol{X}_j} A_j, \\ &\text{s.t.} \quad \| A_i \|_2 = 1, \ i = 1, 2, \ldots, m, \end{aligned} \tag{4}$$

where $\| \cdot \|_2$ denotes the L_2 norm and $\Sigma_{\boldsymbol{X}_i \boldsymbol{X}_j}$ is the correlation matrix between two variable groups $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. For clarity, we denote $\boldsymbol{A} = \{A_1, \ldots, A_m\}$ as the parameters to be optimized.

The intuition behind using MCCA to learn the representation is that we want the representations to distill the causal relations through maximizing the correlation (a relaxation of causality) between two groups. Intuitively, if group Y_j is Granger causally related to group Y_i, then these two groups are

supposed to be highly correlated. Note that if two groups Y_i and Y_j are affected by the same common causes but no directly connected in the graph, there will be a spurious causal relationship between Y_i and Y_j. In order to avoid this kind of spurious causal relationship, we introduce the causal structure as a mask to guide the model to focus on the correlation between causal pairs. Then, we further have the following objective function:

$$\begin{aligned} &A_1^*, A_2^*, \ldots, A_m^* = \arg\max \sum_{i=1}^{m} \sum_{j=i+1}^{m} W_{i,j} A_i^{\mathrm{T}} \Sigma_{\boldsymbol{X}_i \boldsymbol{X}_j} A_j, \\ &\text{s.t. } \parallel A_i \parallel_2 = 1,\ i = 1, 2, \ldots, m, \end{aligned} \tag{5}$$

where $W_{i,j}$ is a binary variable, implying whether X_i is directly connected with X_j.

One may be concerned with the linearity of the presentation, the availability of the causal structure, the identifiability of the representation, and so on issues. Regarding the linearity, the first reason is that we mainly follow the existing work [3,9,20] which simply uses the mean of the variables in a group as the representation; the second reason is that we want to use a relatively simple but explainable method to learn the representation; the third reason is that an overly complex nonlinear model (e.g., long short term memory (LSTM)) may lead to unreasonable high correlations between any two groups in the maximizing of the correlation. Regarding the availability of the causal structure and the identifiability, we tend to take the representation learning as a part of the overall model and will be jointly learned with the causal structure in Subsection 4.2.

4.2 Granger causal structure learning among representations

To learn the representations, we need to discover the Granger causal relationships among them. With the data generating process assumption, for each causal representation Y_i at time t, denoted as $Y_i^{(t)}$, $i \in \{1, \ldots, m\}$, we can construct a Granger causal model for representations as

$$Y_i^{(t)} = g_i(\boldsymbol{PA}_i^{(t)}) + E_i^{(t)} (i = 1, 2, \ldots, m), \tag{6}$$

where g_i is a continuous function that specifies how the variable sets $\boldsymbol{PA}_i^{(t)}$ are mapped to $Y_i^{(t)}$, and $E_i^{(t)}$ is the noise term. If Y_j is not the Granger cause of Y_i, then the function g_i does not depend on Y_j. Based on this, we can derive the definition of Granger non-causality as the Definition 3.

Definition 3. Let $\boldsymbol{Y} = Y_1, Y_2, \ldots, Y_m$ denote m-th time series, which are generated by (6). Then Y_j is not the Granger cause of Y_i if for all $(Y_1, Y_2, \ldots, Y_m)$ and all $Y_j' \neq Y_j$,

$$g_i(Y_1, Y_2, \ldots, Y_j, \ldots, Y_m) = g_i(Y_1, Y_2, \ldots, Y_j', \ldots, Y_m), \tag{7}$$

that is, g_i is invariant to Y_j.

To explicitly formalize the Granger (non-)causal relationship among representations, the model (6) is formalized as

$$\begin{aligned} Y_i^{(t)} &= g_i(q_i(\boldsymbol{Y}^{(<t)})) + E_i^{(t)} \\ &= g_i(q_{i1}(Y_1^{(<t)}), \ldots, q_{im}(Y_m^{(<t)})) + E_i^{(t)}, \end{aligned} \tag{8}$$

$$q_{ij}(Y_j^{(<t)}) = \begin{cases} Y_j^{(<t)}, & \text{if } Y_j \text{ is the cause of } Y_i, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where q_{ij} is the indicator function that indicates if Y_j is the Granger cause of Y_i, $j \in \{1, \ldots, m\}$, $Y_j^{(<t)} = (\ldots, Y_j^{(t-2)}, Y_j^{(t-1)})$ denotes the time series of Y_j up to time t, and $\boldsymbol{Y}^{(<t)} = ((Y_1^{(<t)})^{\mathrm{T}}, \ldots, (Y_m^{(<t)})^{\mathrm{T}})^{\mathrm{T}}$.

Considering the data is temporal and the LSTM [22] compresses the entire past time series using the recursive updates of the hidden state, we utilize the LSTM to model the dependence on the past time series $\boldsymbol{Y}^{(<t)}$, similar to [7].

To make the model clear, we consider the indicator function q_{ij} to be in the linear form, but note that the indicator function can be generalized to a more complicated form. In particular, we have

$$q_{ij}(Y_j^{(<t)}) = W_{i,:j} Y_j^{(<t)}, \tag{10}$$

where $W_{i,:j}$ is the j-th column of $\boldsymbol{W}_i$.

Based on the definition of Granger non-causality and the above formalization, we can derive Proposition 1.

Proposition 1. Assume that Granger causal representations $\boldsymbol{Y}$ are generated by (8). Give the Granger causal structure $\mathcal{G}(\boldsymbol{Y}, \boldsymbol{E_Y})$, where $\boldsymbol{E_Y}$ denotes all Granger causal edges in $\mathcal{G}$. Then $Y_j \to Y_i \notin \boldsymbol{E_Y}$, if and only if $W_{i,:j} = 0$.

Proof. (1) "If" part. If $W_{i,:j} = 0$, then for any $Y_j \neq Y_j'$ for all time series, we have

$$\begin{aligned} q_{ij}(Y_j^{(<t)}) &= W_{i,:j} Y_j^{(<t)} = 0, \\ q_{ij}(Y_j'^{(<t)}) &= W_{i,:j} Y_j'^{(<t)} = 0. \end{aligned} \tag{11}$$

Therefore, according to (9), Y_j is not the Granger cause of Y_i.

(2) "Only if" part. If $Y_j \to Y_i \notin \boldsymbol{E_Y}$, then according to Definition 3, there must be $Y_i \neq Y_i'$ for all time series, and

$$g_i(Y_1, Y_2, \ldots, Y_j, \ldots, Y_m) = g_i(Y_1, Y_2, \ldots, Y_j', \ldots, Y_m). \tag{12}$$

Then according to (8), we have

$$\begin{aligned} &g_i(q_{i,1}(Y_1^{(<t)}), q_{i,2}(Y_2^{(<t)}), \ldots, q_{ij}(Y_j^{(<t)}), \ldots, q_{i,m}(Y_m^{(<t)})) \\ &= g_i(q_{i,1}(Y_1^{(<t)}), q_{i,2}(Y_2^{(<t)}), \ldots, q_{ij}(Y_j'^{(<t)}), \ldots, q_{i,m}(Y_m^{(<t)})). \end{aligned} \tag{13}$$

According to (10), we can obtain

$$\begin{aligned} q_{ij}(Y_j^{(<t)}) &= q_{ij}(Y_j'^{(<t)}), \\ W_{i,:j} Y_j^{(<t)} &= W_{i,:j} Y_j'^{(<t)}. \end{aligned} \tag{14}$$

Thus there must be $W_{i,:j} = 0$.

Under Proposition 1, we can minimize the prediction error of the $Y_i^{(t)}$ by applying the LSTM model. Using the least squares method to estimate the prediction error of the $Y_i^{(t)}$, we can derive the following objective function:

$$\min_{\boldsymbol{\theta}_i} \sum_{t=2}^{T} (Y_i^{(t)} - g_i(q_i(\boldsymbol{Y}^{(<t)})))^2 + \lambda \sum_{j=1}^{m} \parallel W_{i,:j} \parallel_1, \tag{15}$$

where $\boldsymbol{\theta}_i$ denotes the trainable parameter set $\{\boldsymbol{W}_i, \boldsymbol{U}_i, W_i^{\text{out}}\}$, λ is a hyper-parameter, and $\parallel W_{i,:j} \parallel_1$ is a lasso penalty across columns of $\boldsymbol{W}_i$ that is added to select for which variables Granger cause $Y_i^{(t)}$. Many columns of $\boldsymbol{W}_i$ will be zero with a large enough λ, leading to a sparse Granger causal graph.

For all m variables, Eq. (15) can be extended as

$$\begin{aligned} &\sum_{i=1}^{m} \left(\min_{\boldsymbol{\theta}_i} \sum_{t=2}^{T} (Y_i^{(t)} - g_i(q_i(\boldsymbol{Y}^{(<t)})))^2 + \lambda \sum_{j=1}^{m} \parallel W_{i,:j} \parallel_1 \right) \\ &= \min_{\boldsymbol{\theta}} \sum_{t=2}^{T} \sum_{i=1}^{m} (Y_i^{(t)} - g_i(q_i(\boldsymbol{Y}^{(<t)})))^2 + \lambda \sum_{i=1}^{m} \sum_{j=1}^{m} \parallel W_{i,:j} \parallel_1, \end{aligned} \tag{16}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ are parameters to be optimized.

In summary, for $i \in \{1, \ldots, m\}$, the Granger causes of variables Y_i can be interpreted by the parameters $W_{i,:j}$, and the numerical values of $W_{i,:j}$ can be derived by solving the optimization problem, Eq. (16). The architecture of this solution for Granger causal discovery is illustrated as Figure 2(b).

4.3 Model summarization

Till this end, we can combine the Granger causal model and its LSTM representation discussed in Subsection 4.2, with the Granger causal representation introduced in Subsection 4.1. We propose an HGCRM for learning the representation and Granger causal structure for the Group of time series, which is outlined in Figure 2. Specifically, the HGCRM is a two-level model. The first level is identifying group

variables and learning representations, and the second level is identifying the Granger causal relationships among representations. Given multiple groups of time series, the proposed model can be formalized as the following optimization setting:

$$\begin{aligned}&\min_{\boldsymbol{\theta}} \sum_{t=2}^{T}\sum_{i=1}^{m}(Y_i^{(t)} - g_i(q_i(\boldsymbol{Y}^{(<t)})))^2 + \lambda\sum_{i=1}^{m}\sum_{j=1}^{m} \parallel W_{i,:j} \parallel_1 - \sum_{i=1}^{m}\sum_{j=i+1}^{m} W_{i,:j} A_i^{\mathrm{T}} \Sigma_{\boldsymbol{X}_i \boldsymbol{X}_j} A_j,\\ &\text{s.t.} \quad \parallel A_i \parallel_2 = 1,\ i = 1, 2, \ldots, m,\end{aligned} \tag{17}$$

where $\boldsymbol{\theta}, \boldsymbol{A}$ are the decision variables and $Y_i^{(t)} = A_i^{\mathrm{T}} X_i^{(t)}$.

With the formalization of the model, it is worth noting that:

(1) Minimizing the prediction error on $Y_i^{(t)}$ by applying the Granger causal model with LSTM presentation, and finding the optimal vector A_i^*, can be optimized simultaneously;

(2) When the number of measured variables in each group is equal to 1, we have $A_i = 1$ to satisfy the constraint and hence $Y_i^{(t)} = X_i^{(t)}$. Meaning that the problem formulation we proposed in (17), is general enough that it contains which de facto falls back to the simple and most common case;

(3) The above optimization problem can be solved by numerical-based approaches, e.g., the Lagrangian method, which will be discussed shortly.

Once we solve the optimization setting in (17), we can obtain the Granger causes of each group and the Granger causal graph among groups.

5 Augmented-Lagrangian-based model training

The objective function listed in (17) can be solved by applying the augmented Lagrangian method, which redefines the original problem as follows:

$$\begin{aligned}L(\boldsymbol{\theta}, \boldsymbol{A}, \alpha) =& \sum_{t=2}^{T}\sum_{i=1}^{m}(Y_i^{(t)} - g_i(q_i(\boldsymbol{Y}^{(<t)})))^2\\ &+ \lambda\sum_{i=1}^{m}\sum_{j=1}^{m} \parallel W_{i,:j} \parallel_1 - \sum_{i=1}^{m}\sum_{j=i+1}^{m} W_{i,:j} A_i^{\mathrm{T}} \Sigma_{\boldsymbol{X}_i \boldsymbol{X}_j} A_j\\ &+ \sum_{i=1}^{m} \alpha_i h(A_i) + \frac{\mu}{2}\sum_{i=1}^{m} |h(A_i)|^2,\end{aligned} \tag{18}$$

where $h(A_i) := \parallel A_i \parallel_2 - 1$, α_i is the Lagrange multiplier, and $\mu > 0$ is the penalty parameter.

We apply the following update rules for the $(n+1)$-th step to search for the optimal solution to (18),

$$\boldsymbol{A}^{n+1} = \arg\min_{\boldsymbol{A}} L^n(\boldsymbol{\theta}, \boldsymbol{A}, \alpha), \tag{19}$$

$$\boldsymbol{\theta}^{n+1} = \arg\min_{\boldsymbol{\theta}} L^n(\boldsymbol{\theta}, \boldsymbol{A}, \alpha), \tag{20}$$

$$\alpha_i^{n+1} = \alpha_i^n + \mu^n h(A_i^{n+1}), \quad \text{for } i = 1, 2, \ldots, m, \tag{21}$$

$$\mu^{n+1} = \begin{cases} \beta\mu^n, & \text{if } \sum_{i=1}^{m} |h(A_i^{n+1})| \geqslant \gamma \sum_{i=1}^{m} |h(A_i)|,\\ \mu^n, & \text{otherwise,}\end{cases} \tag{22}$$

where $\beta > 1$ and $\gamma < 1$ are two tuning parameters, similar to [23]. Eq. (19) is a first-order differentiable and can be solved by applying a gradient descent method. We use proximal gradient descent to deal with (20) for the sake of the lasso penalty, which can ultimately lead to exact zero by Proximal optimization [24] and it is critical for interpreting causality in our context.

Based on the above update rules, we can simultaneously train $\boldsymbol{A}$ and $\boldsymbol{\theta}$. In detail, we restrict the observed data to the Granger causal representation by $\boldsymbol{A}$, and learn the parameter $\boldsymbol{\theta}$. After obtaining the parameters $\{\boldsymbol{\theta}^n, \boldsymbol{A}^n, \alpha^n\}$, we calculate $L^n(\boldsymbol{\theta}, \boldsymbol{A}, \alpha)$ as (18). Then those parameters are updated according to (19)–(22), and iterate the above two steps until convergence.

6 Experiments

In order to evaluate the effectiveness of our proposed approach, we compare our method together with the baseline methods on synthetic data and real-world data. In all the experimental settings, multivariate Granger causality (MVGC) [25], PCMCI [15], DYNOTEARS [17], and TiMINo [5] are used as the baseline methods, and the implementations of MVGC[1)] and TiMINo[2)] are publicly accessible.

6.1 Synthetic data

In this subsection, we conduct a set of experiments to evaluate the performance of our method on synthetic data. Because these baseline methods are not suitable for estimating directly the causal structure from multigroup time series, we make appropriate modifications to these methods accordingly while ensuring that the basic ideas behind them remain unchanged. Specificity, all modified methods perform two stages separately, which first learn the representation and then estimate the causal structure. We use "Mean-" to denote the method that applies the baseline method to the representations that are token mean values of the measured variables; and use "MCCA-" to denote the method that applies the baseline to the canonical variable for each variable group learned by MCCA. All the baseline methods are implemented based on their original source codes with the default parameters.

To make sure that the simulated multigroup time series are Granger causally connected at the group granularity, we first generate $\boldsymbol{Y}$ with the Lorenz-96 model [26]. The Lorenz-96 model is a typical nonlinear model for capturing the essence of a problem, which can be used to represent the causal dynamic systems. A m-dimensional Lorenz model is

$$\frac{\mathrm{d}Y_i^{(t)}}{\mathrm{d}t} = (Y_{i+1} - Y_{i-2})Y_{i-1} - Y_i + F, \tag{23}$$

where $Y_{-1} = Y_{m-1}, Y_0 = Y_m, Y_{m+1} = Y_1$ and F is a forcing constant (we set $F = 10$) which determines the level of nonlinearity and chaos of the time series. The above generating process results in a multivariate, nonlinear time series with sparse causal connections, similar to [27]. Then we decompose $Y_i \in \mathbb{R}^{1\times T}$ into p_i time series with the following equation:

$$\boldsymbol{X}_i = B_i Y_i, \text{ for } i = 1, 2, \ldots, m, \tag{24}$$

where B_i is a $p_i \times 1$ dimensional weighting vector.

We generate simulated time series data with the varying number of groups (denoted as m), the average number of variables in each group (i.e., $\frac{1}{m}\sum_{i=1}^{m} p_i$, and p_i is set randomly), the length of time series, respectively. In detail, we conduct experiments on three cases as follows.

(1) The number of groups $= \{8, 10, 12, 14, 16\}$, the average number of variables in a group is 4 and the length of the time series is 1500.

(2) The average number of measured variables in each group $= \{2, 3, 4, 5, 6\}$, the number of groups is 12 and the length of the time series is 1500.

(3) The length of time series $= \{500, 1000, 1500, 2000, 2500\}$, the average number of variables in each group is 4 and the number of groups is 12.

We count the number of true positives (TP, the number of edges in the estimated graph also present in the ground-truth), false positives (FP, the number of edges in the estimated graph not present in the ground-truth), and false negatives (FN, the number of edges not in the estimated graph but present in the ground-truth), and report the Precision $= \frac{\mathrm{TP}}{\mathrm{TP+FP}}$, Recall $= \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, and F1 $= \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision+Recall}}$ as the evaluation indicators for all algorithms, averaged over 10 times simulation results for each case.

For the proposed model, we simply set $\beta = 1.1$, $\gamma = 0.9$ and only fine-tune the hyper-parameter λ, to make sure that the lasso penalty is not too small in different cases. After we run Mean-TiMINo and MCCA-TiMINo on the simulated data, both Mean-TiMINo and MCCA-TiMINo do not work and return no result. Because the TiMINo algorithm needs a large sample size to perform the regression and independence test, which are not satisfied in the experiments. In all simulations, both Mean-TiMINo and MCCA-TiMINo reject all independence tests with the p-value threshold 0.05 or 0.01, because the sample size is not large enough and the regression method may not remove all influence on the assumed leaf variable. Thus, we only show the experimental results of our method, Mean-GC, MCCA-GC, Mean-PCMCI, MCCA-PCMCI, Mean-DYNOTEARS, and MCCA-DYNOTEARS as given in Figures 3–5.

1) Multivariate Granger causality. http://users.sussex.ac.uk/~lionelb/MVGC.

2) TiMINo. http://web.math.ku.dk/~peters/code.html.

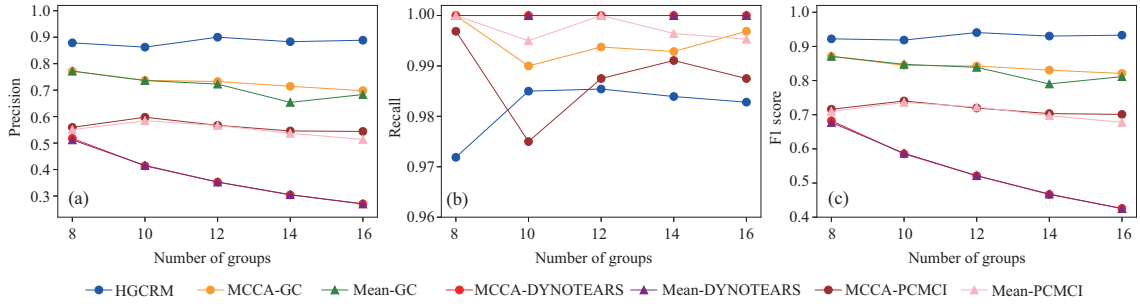

Figure 3 (Color online) Sensitivity to the number of groups. (a) Precision, (b) recall, and (c) F1 score of the recovered causal structure.

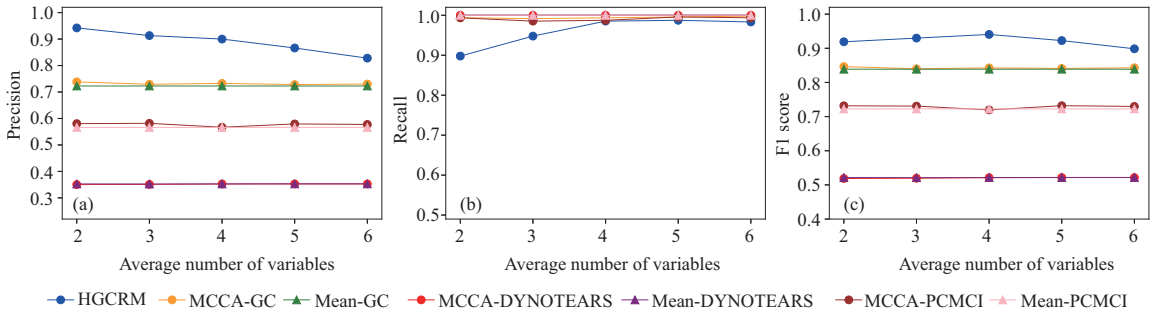

Figure 4 (Color online) Sensitivity to the average number of variables in each group. (a) Precision, (b) recall, and (c) F1 score of the recovered causal structure.

Sensitivity to the number of groups. As shown in Figure 3, both the precision and F1 scores of our method are the highest among the compared methods and all methods have almost the same recall rate. Because the learned causal representations help reduce the redundancy edges among observed variables and the powerful fitting ability of the neural network, our method removes most of the false edges and thus achieves higher precision and F1 scores than others. Figure 3 further illustrates that our method performs more robustly and has a steady F1 score when the number of sets increases, while the F1 scores of other compared methods decrease.

Sensitivity to the average number of variables in each group. As shown in Figure 4, the precision of our method is higher than the others in all cases. When the average number of variables in each group increases, the precision of our method decreases slightly. This is because the number of the measured variables increases, the complexity of learning the causal structure becomes higher. On the contrary, the precision of MCCA-GC and Mean-GC stay close to 0.72 in all cases. It illustrates that the results of MCCA-GC and Mean-GC contain some false causal relationships. These two methods obtain the group variables first, so they, in fact, learn the causal structure from the fixed group variables while the group variables may be spurious correlated. And our method obtains the canonical variables dynamically and has a better chance to learn the correct causal relationships. The recall of all methods is very high and close to 1. Comparing the F1 score of these methods, we can find that our method obtains the best performance.

Sensitivity to the length of time series. Figure 5 illustrates the performance of the algorithms with a different length of time series. As shown in Figure 5, the precision and F1 scores of our method not only grow much faster than the compared methods but also get higher values when the length of the time series increases in most cases. The only exception happens when the length of the time series is 500. This is because our method utilizes the recurrent neural network (RNN)-based approach to search for the nonlinear causal relation of variables and inherently requires more data to correctly estimate the parameters of the model. In real-world scenarios, the sample size is usually larger than 500, therefore our method can be applied to obtain better results.

Ablation study. To measure the effectiveness of each component of our proposed method, we introduce three variants of our model as follows.

• **MCCA-HGCRM.** We remove the Granger causal constraint on learning the representations and

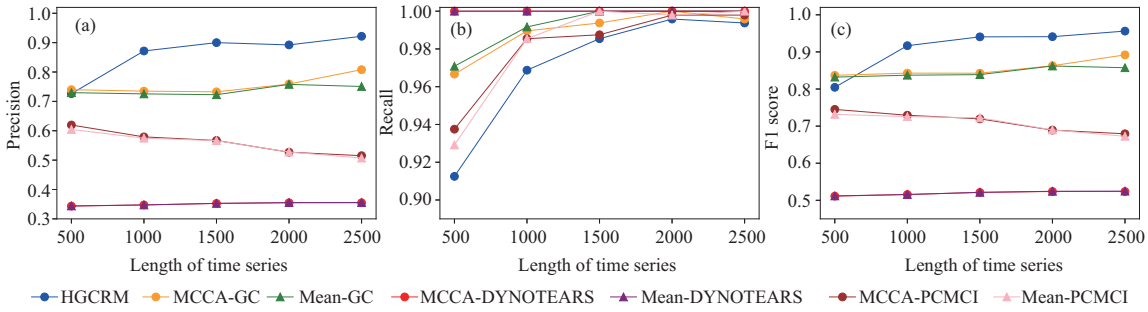

Figure 5 (Color online) Sensitivity to the length of time series. (a) Precision, (b) recall, and (c) F1 score of the recovered causal structure.

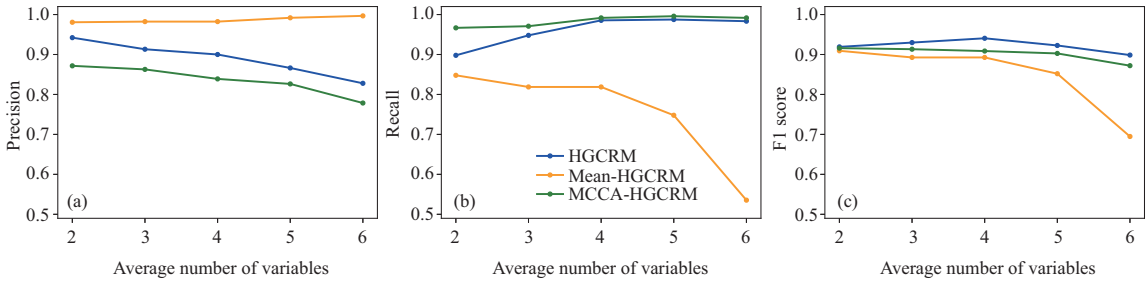

Figure 6 (Color online) Experimental results in the different average number of variables. (a) Precision, (b) recall, and (c) F1 score of the recovered causal structure.

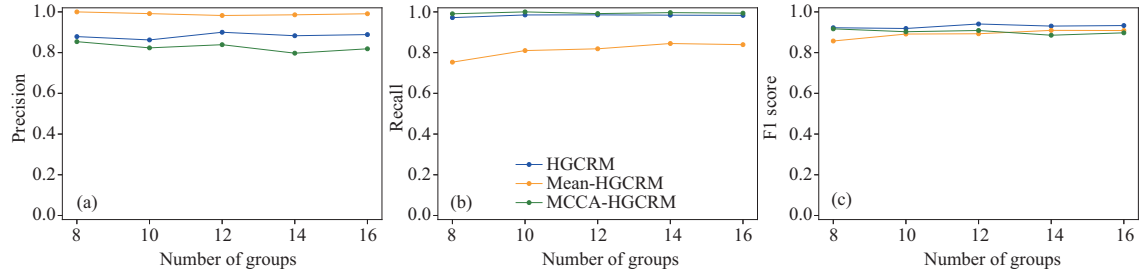

Figure 7 (Color online) Experimental results in different number of groups. (a) Precision, (b) recall, and (c) F1 score of the recovered causal structure.

change the objective function as

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{A}, \alpha) = & \sum_{t=2}^{T} \sum_{i=1}^{m} (Y_i^{(t)} - g_i(q_i(\boldsymbol{Y}^{(<t)})))^2 \\ & + \lambda \sum_{i=1}^{m} \sum_{j=1}^{m} \parallel W_{i,:j} \parallel_1 - \sum_{i=1}^{m} \sum_{j=i+1}^{m} A_i^{\mathrm{T}} \Sigma_{\boldsymbol{X}_i \boldsymbol{X}_j} A_j \\ & + \sum_{i=1}^{m} \alpha_i h(A_i) + \frac{\mu}{2} \sum_{i=1}^{m} |h(A_i)|^2. \end{aligned} \tag{25}$$

Then, we obtain the result by minimizing (25).

• **Mean-HGCRM.** In this variant, we change the learning representation method with MCCA to taking the average values of measured variables as the representations. Similar to the modification of the baselines, Mean-HGCRM is a two-stage method, which first learns the representations first by taking mean values and then estimates the Granger causality among representations.

The results of HGCRM and the above variants are given in Figures 6–8, From the results, the performance of HGCRM is best in recall and F1 score, which shows that both the Granger causal relationships weight restriction and the MCCA idea help to learn the representations and the Granger causal relationships among representations. Comparing the results of Mean-HGCRM with that of two MCCA-related

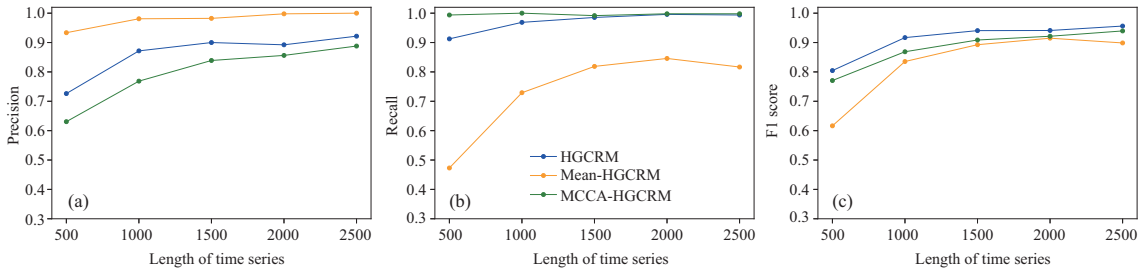

Figure 8 (Color online) Experimental results in different length of time series. (a) Precision, (b) recall, and (c) F1 score of the recovered causal structure.

methods (including MCCA-HGCRM and HGCRM), the recalls show that using the mean value as the representation loses information on recovering the Granger causal relationships. Although the precision of Mean-HGCRM is higher than the other two methods, the reason is that the learned causal relationships by Mean-HGCRM are mostly their own influence relationships, which shows that the causal relationships between different representations are lost.

6.2 Real-world dataset

To further assess the performance of our model, we carry out experiments for performance comparison among these methods on the real resting-state functional magnetic resonance imaging (rs-fMRI) dataset, which is a subset of the enhanced NKI Rockland sample [28]. In this dataset, the BOLD signals are measured and collected at the voxel level from fMRI. Thus, the voxel-level time series data are taken as input and different ROIs are regarded as different groups. Each voxel-level time series belongs to one ROI, which is determined by the anatomical parcellation.

In this experiment, resting-state fMRI scans (TR = 645 ms) are preprocessed and projected onto the Freesurfer fsaverage5 template [29, 30] using a python package NiLearn [31]. The Destrieux parcellation [32] in fsaverage5 space as distributed with Freesurfer is used to select the seed regions. We consider the following seven ROIs: posterior cingulate cortex (PCC), anterior cingulate cortex, middle temporal gyrus and angular gyrus in the left hemisphere (LACC, LMTG, LAG, respectively), and anterior cingulate cortex, middle temporal gyrus, and angular gyrus in the right hemisphere (RACC, RMTG, RAG, respectively). These regions are commonly studied and some of them are correlated during the resting state [33–35].

In short, this dataset consists of 1258 voxels (measured time series) comprising seven different ROIs: PCC (116 voxels), LACC (167 voxels), LMTG (183 voxels), LAG (171 voxels), RACC (191 voxels), RMTG (188 voxels) and RAG (242 voxels) with the length of time series 895 and the dataset is public available[3)].

In practice, most ROI-based Granger causal connectivity studies [3, 20] are based on the baseline algorithm Mean-GC, which first takes average values of the voxel time series within each ROI as the representation (which is an individual time series in this context) of the corresponding ROI, and then applies Granger causality to the ROI level time series to obtain the causal connections of ROIs. Therefore, these two methods, Mean-GC and MCCA-GC, are typical representatives for performance comparison. Besides, we also modified DYNOTEARS and PCMCI algorithms with representations obtained by mean values or MCCA as the baseline methods. The results estimated by our method and modified baseline methods are shown in Figure 9.

According to Figure 9(a), our method outputs a compact, clear, and interpretable result that the PCC is the main cortical hub, and PCC is the cause of AG, MTG, and LACC. Comparing with the background knowledge, we find that this result is consistent with previous studies where the work on [35] found a striking positive correlation between the ACC and PCC, and the work on [34] found the PCC was positively correlated with the RAG and the bilateral MTG. With the sparse constraint, the results of MCCA-DYNOTEARS and Mean-DYNOTEARS are slightly sparse. In contrast, as illustrated in Figures 9(b), (f), and (g), more edges are found and connected massively which is difficult to extract the main causal connectivities. In our causal graph, the bi-directional edges like LAG $\leftrightarrow$ LMTG mean the LAG affects the LMTG with a time lag and the LMTG also affects the LAG with a time lag. Figure 9(c)

3) An example to process and acquire the dataset. http://nilearn.github.io/auto_examples/01_plotting/plot_surf_stat_map.html.

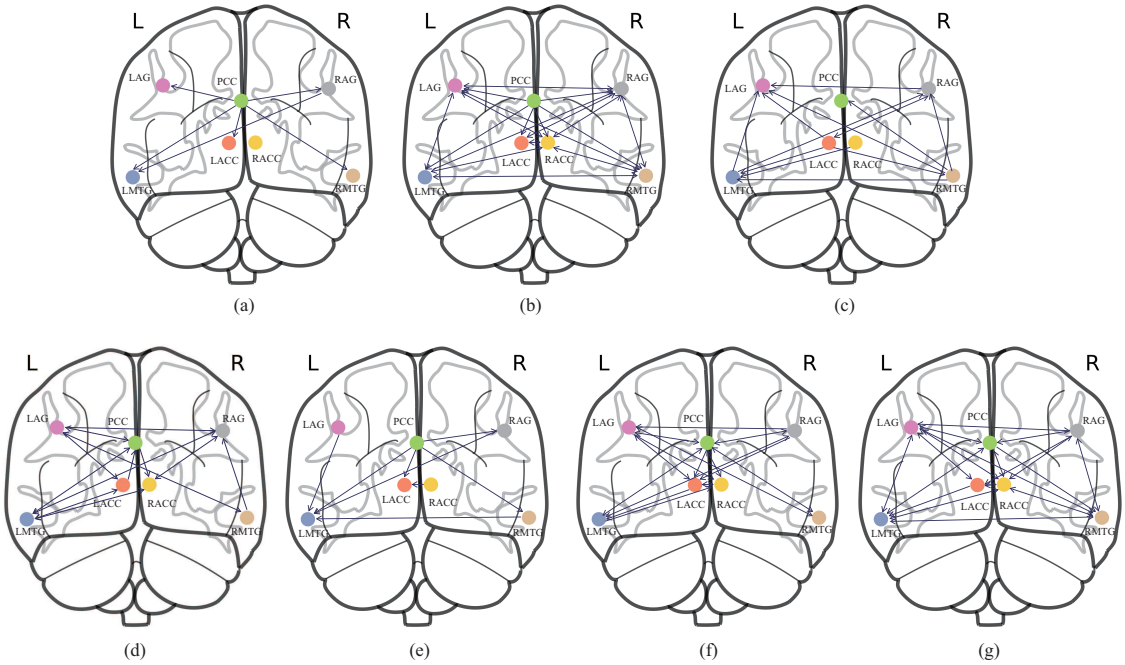

Figure 9 (Color online) Causal graphs of 7 selected ROIs estimated by our method and baseline methods from real-world fMRI dataset. (a) HGCRM; (b) MCCA-GC; (c) Mean-GC; (d) MCCA-DYNOTEARS; (e) Mean-DYNOTEARS; (f) MCCA-PCMCI; (g) Mean-PCMCI.

shows the result of the Mean-GC and edges like LMTG $\rightarrow$ LAG is consistent with [36], but the PCC is only connected with the RMTG and becomes less significant than other brain regions. The PCC has been suggested to be a cortical hub during the resting state [35] while the Mean-GC failed to report this important result. Moreover, these results also confirm that our method successfully removes most of the false edges and further explain why our method earns higher precision in the synthetic datasets.

7 Conclusion

Learning the Granger causal structure among groups of variables from multigroup time series is an important problem and has many practical applications. In this paper, we explore a hierarchical Granger causal model for Granger causal discovery from multiple groups of time series. First, we discuss a new Granger causal representation for groups of time series. Second, we introduce a nonlinear Granger causal model to estimate the Granger causal relationships among the representations. Third, we propose to formulate an optimization objective by simultaneously considering the nonlinear Granger causal model, LSTM representation, and MCCA approach. Next, we show that this optimization problem can be solved using the augmented Lagrangian method. Finally, we conduct several experiments based on synthetic and real-world datasets for performance evaluation. The experimental results show that strengthening the correlations of variables may introduce false edges; however, our method can remove these false edges and discover meaningful causal relations with the highest precision among all baseline methods.

As a final note, in this study, we assume that the groups of the variables are given. If this assumption is invalid, we must consider how to cluster variables into different groups and then learn the Granger causal relationships among variable groups, which will be considered in our future work.

Acknowledgements This work was supported in part by National Key R&D Program of China (Grant No. 2021ZD0111501), National Science Fund for Excellent Young Scholars (Grant No. 62122022), National Natural Science Foundation of China (Grant Nos. 61876043, 61976052, 62206064), and the Major Key Project of PCL (Grant No. PCL2021A12).

References

1 Chen W, Cai R C, Hao Z F, et al. Mining hidden non-redundant causal relationships in online social networks. Neural Comput Applic, 2020, 32: 6913–6923
2 Cai R C, Zhang Z J, Hao Z F. Causal gene identification using combinatorial V-structure search. Neural Netw, 2013, 43: 63–71

3 Sanchez-Romero R, Ramsey J D, Zhang K, et al. Estimating feedforward and feedback effective connections from fMRI time series: assessments of statistical methods. Netw Neurosci, 2019, 3: 274–306
4 Hyvärinen A, Zhang K, Shimizu S, et al. Estimation of a structural vector autoregression model using non-gaussianity. J Mach Learn Res, 2010, 11: 1709–1731
5 Peters J, Janzing D, Schölkopf B. Causal inference on time series using restricted structural equation models. In: Proceedings of Advances in Neural Information Processing Systems, 2013. 154–162
6 Granger C W J. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 1969, 37: 424–438
7 Tank A, Covert I, Foti N, et al. Neural Granger causality. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 4267–4279
8 Löwe S, Madras D, Zemel R, et al. Amortized causal discovery: learning to infer causal graphs from time-series data. In: Proceedings of Conference on Causal Learning and Reasoning, 2022. 509–525
9 Huang B W, Zhang K, Sanchez-Romero R, et al. Diagnosis of autism spectrum disorder by causal influence strength learned from resting-state fMRI data. 2019. ArXiv:1902.10073
10 Entner D, Hoyer P O. Estimating a causal order among groups of variables in linear models. In: Proceedings of International Conference on Artificial Neural Networks, 2012. 84–91
11 Parviainen P, Kaski S. Learning structures of Bayesian networks for variable groups. Int J Approximate Reason, 2017, 88: 110–127
12 Spirtes P, Glymour C N, Scheines R, et al. Causation, Prediction, and Search. Cambridge: MIT Press, 2000
13 Zhang K, Huang B W, Schölkopf B, et al. Towards robust and specific causal discovery from FMRI. 2015. ArXiv:1509.08056
14 Chu T J, Glymour C. Search for additive nonlinear time series causal models. J Mach Learn Res, 2008, 9: 967–991
15 Runge J, Nowack P, Kretschmer M, et al. Detecting and quantifying causal associations in large nonlinear time series datasets. Sci Adv, 2019, 5: 4996
16 Ashrafulla S, Haldar J P, Joshi A A, et al. Canonical Granger causality between regions of interest. NeuroImage, 2013, 83: 189–199
17 Pamfil R, Sriwattanaworachai N, Desai S, et al. Dynotears: structure learning from time-series data. In: Proceedings of International Conference on Artificial Intelligence and Statistics, 2020. 1595–1605
18 Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet, 2019, 10: 524
19 Scholkopf B, Locatello F, Bauer S, et al. Toward causal representation learning. Proc IEEE, 2021, 109: 612–634
20 Marinazzo D, Liao W, Chen H F, et al. Nonlinear connectivity by Granger causality. NeuroImage, 2011, 58: 330–338
21 Li Y O, Adali T, Wang W, et al. Joint blind source separation by multiset canonical correlation analysis. IEEE Trans Signal Process, 2009, 57: 3918–3929
22 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780
23 Ng I, Zhu S Y, Chen Z T, et al. A graph autoencoder approach to causal structure learning. 2019. ArXiv:1911.07420
24 Gong P H, Zhang C S, Lu Z S, et al. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In: Proceedings of International Conference on Machine Learning, 2013. 37–45
25 Barnett L, Seth A K. The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. J Neurosci Methods, 2014, 223: 50–68
26 Karimi A, Paul M R. Extensive chaos in the Lorenz-96 model. Chaos-An Interdisc J Nonlinear Sci, 2010, 20: 043105
27 Tank A, Cover I, Foti N J, et al. An interpretable and sparse neural network model for nonlinear Granger causality discovery. 2017. ArXiv:1711.08160
28 Nooner K B, Colcombe S J, Tobe R H, et al. The NKI-rockland sample: a model for accelerating the pace of discovery science in psychiatry. Front Neurosci, 2012, 6: 152
29 Dale A M, Fischl B, Sereno M I. Cortical surface-based analysis: I. segmentation and surface reconstruction. NeuroImage, 1999, 9: 179–194
30 Fischl B, Sereno M I, Dale A M. Cortical surface-based analysis: II. ination, attening, and a surface-based coordinate system. NeuroImage, 1999, 9: 195–207
31 Abraham A, Pedregosa F, Eickenberg M, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform, 2014, 8: 14
32 Destrieux C, Fischl B, Dale A, et al. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. NeuroImage, 2010, 53: 1–15
33 Garza-Villarreal E A, Jiang Z, Vuust P, et al. Music reduces pain and increases resting state fMRI BOLD signal amplitude in the left angular gyrus in fibromyalgia patients. Front Psychol, 2015, 6: 1051
34 Tan X, Liang Y, Zeng H, et al. Altered functional connectivity of the posterior cingulate cortex in type 2 diabetes with cognitive impairment. Brain Imag Behav, 2019, 13: 1699–1707
35 Cao W F, Luo C, Zhu B, et al. Resting-state functional connectivity in anterior cingulate cortex in normal aging. Front Aging Neurosci, 2014, 6: 280
36 Guo W B, Liu F, Xiao C Q, et al. Increased causal connectivity related to anatomical alterations as potential endophenotypes for schizophrenia. Medicine, 2015, 94: 1493