Unified Model Selection Approach Based on Minimum Description Length Principle in Granger Causality Analysis

FEI LI [ID], XUEWEI WANG, QIANG LIN, AND ZHENGHUI HU [ID]
College of Science, Zhejiang University of Technology, Hangzhou 310023, China

Corresponding author: Zhenghui Hu (e-mail: zhenghui@zjut.edu.cn).

ABSTRACT Granger causality analysis (GCA) provides a powerful tool for uncovering the patterns of brain connectivity mechanism using neuroimaging techniques. In this paper, distinct from conventional two-stage GCA, we present a unified model selection approach based on the minimum description length (MDL) principle for GCA in the context of the general regression model paradigm. In comparison with conventional methods, our approach emphasizes that model selection should follow a single mathematical theory during the GCA process. Under this framework, all candidate models within the model space might be compared freely in the context of the code length, without the need for an intermediate model. We illustrated its advantages over conventional two-stage GCA approach in a 3-node network and a 5-node network synthetic experiments. The unified model selection approach was capable of identifying the actual connectivity while avoiding the false influences of noise. More importantly, the proposed approach obtained more consistent results in a challenging fMRI dataset, in which visual/auditory stimulus with the same presentation design gives identical neural correlates of mental calculation, allowing one to evaluate the performance of different GCA methods. Moreover, the proposed approach has potential to accommodate other Granger causality representations in other function space. The comparison between different GC representations in different function spaces can also be naturally deal with in the framework.

INDEX TERMS Code length, Granger causality analysis (GCA), minimum description length (MDL), model selection.

I. INTRODUCTION

Causal connectivity analysis, also called effective connectivity, plays an increasingly important role in brain research using neuroimaging techniques [1]–[6]. It reflects a trend in neuroscience away from focusing on individual brain unit (functional specialization) toward complex neural circuits (functional integration) at different spatial scales [7]–[9], where the integration among specialized areas is mediated by causal connectivity [10]–[15]. Causal connectivity refers to the influence that one neural system exerts over another, either in the absence of identifiable behavioral events or in the context of task performance [16], [17]. It provides important insights into brain organization. Causal connectivity anal-

ysis is based on temporal relations existing between time series recordings of neural activity, which may be obtained by neuroimaging techniques, such as electroencephalography (EEG) [18]–[20], local field potentials (LFP) [20]–[22], magnetoencephalography (MEG) [23]–[25], and functional magnetic resonance imaging (fMRI) [26]–[29] etc.

Granger causality analysis (GCA) is an effective tool for detecting the causal connections that can provide information about the dynamics and directionality of the associations. Causality in the Granger sense is based on the statistical predictability of one time series that derives from knowledge of one or more other time series [30], [31]. For two time series, X and Y, that are stochastic and wide-sense stationary (i.e., constant means and variances) [32], the general idea of Granger causality is that variable Y is said to have a causal influence on variable X when the prediction of variable X

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan [ID].

could be improved by incorporating the history information of variable Y.

In practical applications, the GCA is usually conveyed in the context of linear autoregressive (AR) models of stochastic processes. The AR models with different time lags represent the historical dependence on variables, with or without including other variables, to forecast the dynamics of the variable. The optimal model for predication then is chosen by model selection techniques. The extension of GCA to nonlinear model that capture nonlinear causal relations is also available [33], [34]. And the validity of original GCA can also be affected if the errors after the prediction with the model is not normally distributed [26], [35]. Several test methods of causality in Granger sense for the data with asymptotical noise distribution have been developed that is especially useful for task-related fMRI studies [36], [37]. Asymmetric causality testing recently has also been suggested in order to separate the causal impacts of positive and negative changes [38]. Moreover, Granger causality has been also expressed in other function spaces, e.g. Fourier spaces (frequency domain) [21], [39]–[45], kernel Hilbert spaces [46], [47]. These methods are important in neuroscience studies since the causal influences between neuronal populations are often nonlinear and have complicated statistics due to various sources of uncertainties. On the other hand, the original GCA is limited to investigate causal connectivity between two nodes. There have been numerous efforts on expanding the approach from small network with few nodes to large, complex network [48], [49]. A recent approach presented an iterative scheme that gradually pruned the network by removing indirect connections, considered one at a time, to uncover large network structure with hundreds of nodes [50].

Despite these developments, the basic idea of Granger causality remains unchanged. It generally comprises two stages in this framework: (1) specifying the model order (the number of time lags that is associated with its own history information and external effects), and (2) deciding the optimal model. The order of the predictors is usually determined with the Akaike information criterion (AIC) [51] or the Bayesian information criterion (BIC) [52], whereas the optimal predictor is judged through statistical pairwise comparison [53]–[56]. Specifying model order and selecting the optimal model comply two completely different mathematical theories.

Indeed, from a mathematical perspective, the two stages are all to select the best model capturing essential features under investigation from a number of competing models in terms of given observations. They are essentially the same problem of model selection. Model selection is the most important aspect of inference with causal models and allows one to test different hypotheses by comparing different models in terms of selection criteria. Two different theories applied in the same question might generate different selection criteria and therefore degrade the performance of GCA.

We argue that the two stages in GCA are the problem of model selection, and should follow the same benchmark under one mathematical theory. In this paper, against two-stage selection scheme, we therefore proposed a unify model selection approach for GCA with the minimum description length (MDL) principle that model selection should follow a single mathematical theory during the GCA process. We illustrated the benefits of introducing a unified model selection approach in simulated and real fMRI experiments. Especially, we compared the proposed approach with conventional two-stage GCA one in an empirical fMRI dataset for the validation of causal connectivity analysis.

The rest of the article is organized as follows. Section II introduces the basic GCA concept and discusses the potential problems in the current two-stage GCA scheme and our motivation for introducing the MDL principle in GCA. In Section III, the MDL principle has been illustrated in detail, and the formula of two-part MDL also has been derived with Bernoulli distribution in Markov model class. Immediately, in Section IV, the MDL guided model selection for linear model has been carried out, in time domain and frequency domain respectively. In Section V, We illustrate the advantages of our proposal over conventional two-stage GCA approach in a 3-node network and a 5-node network synthetic experiments. At the same time, the proposed approach obtained more consistent results in a challenge fMRI dataset for causality investigation, mental calculation network under visual and auditory stimulus, respectively. Section VI demonstrates the comparison between conventional two-stage GCA and our proposal from modeling standpoint, and discusses its potential development in a wider Granger causality sense.

II. PROBLEM STATEMENT

Let variables X and Y be two stochastic and stationary time series. Now consider the following pair of (restricted and unrestricted) regression models

$$\begin{cases} X[n] = \sum_{i=1}^{p} \beta_x[i]X[n-i] + u[n] \\ X[n] = \sum_{i=1}^{p} \beta_x[i]X[n-i] + \sum_{i=1}^{q} \beta_y[i]Y[n-i] + v[n], \end{cases}$$
$$(1)$$

where p and q are the model orders (the numbers of time lags) in X and Y, respectively, β_x and β_y are the model coefficients, and u and v are the residual of the models. The order of historical predictor p is usually determined with the AIC [51] or the BIC [52],

$$AIC = -2log(L(\theta)) + 2k$$
$$BIC = -2log(L(\theta)) + klog(n)$$

where n is the sample size, and k is the number of parameters which your model estimates, and θ is the set of all parameters. $L(\theta)$ represents the likelihood of the model tested, given your data, when evaluated at maximum likelihood values of θ.

The conventional GCA requires statistical significance to determine whether the unrestricted model provides a better prediction than the restricted model. The hierarchical

F-statistics, based on the extra sum-of-squares principle [57], can be used to evaluate significant predictability, given as

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-p-q-1)} \quad vs \quad F_0(q, n-p-q-1) \tag{2}$$

where RSS_r and RSS_u represent the sum of squared residuals of restricted model and unrestricted model, respectively, n is the total number of observations to estimate the unrestricted model. The *F*-statistics approximately follows an *F*-distribution with $(q, n-p-q-1)$ degrees of freedom.

If $F > F_0(q, n-p-q-1)$, the variability of the residual of unrestricted model is significantly less than the variability of the residual of restricted model, then there is an improvement in the prediction of *X* due to *Y*, and we refer to this as causal influence from *Y* to *X*. But in the current bivariate model, spurious connections will emerge frequently.

In order to remove spurious connections caused by indirect causalities between nodes, GCA also provides a measure of conditional causal connection by introducing another variable *Z* into Eqs. (1):

$$\begin{cases} X[n] = \sum_{i=1}^{p} \beta_x[i]X[n-i] + \sum_{i=1}^{r} \beta_z[i]Z[n-i] + u[n] \\ X[n] = \sum_{i=1}^{p} \beta_x[i]X[n-i] + \sum_{i=1}^{q} \beta_y[i]Y[n-i] \\ \quad + \sum_{i=1}^{r} \beta_z[i]Z[n-i] + v[n]. \end{cases} \tag{3}$$

Then, the causal influence from *Y* to *X*, conditional on *Z*, is defined as

$$\mathcal{F}_{Y \to X|Z} = \ln \frac{var(u)}{var(v)}. \tag{4}$$

Consider three model spaces $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, where each space comprises three models. Let p_i and q_i, $i = 1, 2, 3$, denote the model orders regarding its endogenous information and the exogenous information from other variables, respectively. And $n \in N, m \in N$,

$$\mathcal{A} = \{a_1 : p_1 = n, q_1 = 0;$$
$$a_2 : p_2 = n+1, q_2 = 0;$$
$$a_3 : p_3 = n+2, q_3 = 0\}$$
$$\mathcal{B} = \{b_1 : p_1 = n, q_1 = m+0;$$
$$b_2 : p_2 = n, q_2 = m+1;$$
$$b_3 : p_3 = n, q_3 = m+2\}$$
$$\mathcal{C} = \{c_1 : p_1 = n, q_1 = m;$$
$$c_2 : p_2 = n+1, q_2 = m;$$
$$c_3 : p_3 = n+1, q_3 = m+1\}.$$

We further assume that a_i, b_i, and c_i, $i = 1, 2, 3$, have the same residual variance. For space $\mathcal{A}$, the models only specify the regression model orders of the endogenous information by AIC/BIC, and then specify the causal effect of endogenous information by *F*-statistics. The models in space $\mathcal{B}$ specify the regression model orders of the exogenous effect by AIC/BIC, then specify the causal effect of exogenous information by

F-statistics. But the models $\mathcal{C}$ specify the regression model orders of endogenous information and exogenous information separately, and specify the causal effect between c_1, c_2 and c_3 by *F*-statistics. In this situation, the model selection in conventional GCA is split into a two-stage scheme, the model orders is determining by AIC/BIC and then the causal effect is quantified by *F*-statistics sequentially. It is clear that the final inference might differ with rules applied, even though three classes are completely equivalent from a model description standpoint.

As stated above, specifying the effects of endogenous and exogenous information in conventional GCA is split into a two-stage scheme. Specifying the regression model orders of historical information, which contains the regression of endogenous information in restricted model and the regression of both endogenous and exogenous information in unrestricted model, is mainly based on AIC/BIC theory separately, then specifying the causal effect of exogenous information by *F*-statistics. However, specifying the two effects is the same kind of problem of model selection from the mathematical perspective. Different theories might generated different benchmarks in two stages for model selection, therefore degrade the performance of GCA.

Aside from theoretical considerations, there are still some issues in the *F*-test itself to be discussed. The American Statistical Association's (ASA) statement on statistical significance and *p*-values has led to a collective rethinking of the entire scientific community, and many scientists believe that the application of *p*-values in current scientific research has been distorted [58]–[61]. This also leads many studies to selectively report results, and the dichotomous *p*-value is arbitrary reductionism for scientific research. We believe that any reasonable study has its implied meaning, whether or not the *p*-value is less than 0.05. We should understand the true meaning of *p*-values, and it should not be over-exaggerated or degraded [62]–[64]. It is just a statistical tool, or just one of the mapping relationships that is purely mathematical. Specifically, pairwise *F*-statistics in conventional GCA arouses several potential problems that might lead to misleading or unreasonable inferences in connectivity analysis. Firstly, the model comparison with *F*-statistics is performed under a specific significance level. The assignment of significance level is a subjective matter in conventional GCA process [65], [66]. A significance level that is too low could cause the false connection noise originated, whereas a significant level that is too high could erase the actual connection. When there is no rule to justify the assignment of the significance level, *F*-statistics will lead to different connectivity results depending on the significance level chosen.

Secondly, the selection results by pairwise *F*-statistics are heavily dependent on the initially selected model and the search path in model space. Consider the collection of three models $\mathcal{M} = \{A, B, C\}$ and let $\mathcal{S}$ denote the residual variance of each model with $S_A = S$, $S_B = S + \Delta S$, and $S_C = S - \Delta S$. We further assume $\frac{I}{2} \leq \Delta S < I$, where I is the interval satisfying statistical significance. The aim of *F*-statistics is
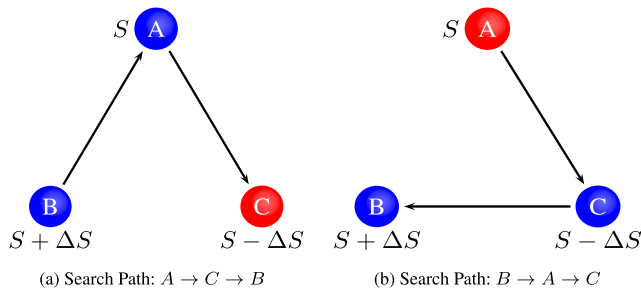
FIGURE 1. Consider a conceivable case involving a collection of three candidate models $\mathcal{M} = \{A, B, C\}$ with residual variances of $S_A = S$, $S_B = S + \Delta S$, and $S_C = S - \Delta S$. Suppose $\frac{I}{2} \leq \Delta S < I$, where I is the interval satisfying statistical significance. It is clear that model C, with the minimum variance, is the optimal model in this case. Following search path $B \to A \to C$ (a) we can arrive at optimal model C, while if we start with A and follow the search path $A \to C \to B$ (b), we reach suboptimal model A. The results using F-statistics for model selection rely on both the search path and the initial model. Comparing transverses within the model space can only partially reduce the risk of model misspecification, and is not always available due to the nested relation between comparable models in F-statistics. Moreover, this strategy also leads to concern about the computational complexity.

to find the model appropriate to the observations from the model space, which is evidently model C in this case. The search path $B \to A \to C$ will arrive at optimal model C, whereas if we start at A and follow the search path $A \to C \to B$, we will reach suboptimal model A, that is, optimal model C is not considered. The distinct results obtained by using F-statistics for model selection along with different search paths and initial models are given explicitly in Fig. 1. In fact, F-statistics can not discriminate the models by residual variance within a specific range relative to the chosen significance level and the noise level. On the other hand, F-statistics uses the extra sum of squares principle to identify a better model. This means that only models with a nested relationship can be compared. Then the competitive models perform pairwise comparison in an indirect way, through *intermediate* model, namely unrestricted model in F-statistics. Therefore, the comparison between any two models is not available in practical applications, and the search path relies on the structure of all candidate models. In such situation the optimal model is not always guaranteed.

The third potential problem relates to computational complexity. Consider the simplest case where conditional causal connections are not taken into account ($q = 1$ and $r = 0$ in Eq. (3)). Suppose that there are m candidate models for any one directional causality in the network with n nodes, the number of F-tests that needs to be performed is $m(m - 1)(n - 1)$, and the total number of comparisons is $mn(m - 1)(n - 1)$. The problem on computational complexity is compounded by conditional GCA, but it still will be intractable while investigating large network [50].

Although different approaches might be applied for model selection in GCA [37], [67], the two stages are kept unchanged. They are generally based on two different mathematical theories in most GCA applications. However, as mentioned above, determining the model orders by AIC/BIC or quantifying the causal effects by F-statistics can

essentially be considered as a generalized model selection problem from a modeling standpoint. And model selection by a single mathematical principle throughout GCA, which could be easily ignored, would determine the final result working pattern of human brain. Thus the causal connectivity obtained by conventional GCA could be misleading, and main reason may be splitting model selection into a two-stage scheme. Since all the issues we discuss can be attributed to model selection problems, then a more practical model selection method need to be enabled.

To keep consistency in model selection for GCA, we proposed a code length guided framework based on MDL principle, which Rissanen first proposed to quantify parsimony of a model [68], to map the two-stage scheme into the same model space. Specifically, our proposal involves constructing a code length guided framework, then the model selection process in GCA can be converted into comparing the code length of each model. That means our proposal incorporated the endogenous and the exogenous information into a unified model selection process, which the information will be quantified by converting into code length to obtain causality. The two-stage scheme based on the two different theories is unified under the single mathematical framework, MDL principle, which guarantees the only benchmark in GCA methodology research.

And above all, the MDL is an information criterion that provides a generic solution for the model selection problem [69]. As a broad principle, the MDL represents a completely different approach for model selection relative to traditional statistical approaches, such as F-statistics and the AIC or BIC. Compared with the AIC/BIC, the use of the MDL does not require any assumptions about the data generation mechanism. In particular, a prior probability distribution does not need to be assigned to each model class. The objective of model selection in the MDL is not to estimate an assumed but unknown distribution, but to find models that more realistically represent the data [70]–[72].

Fundamentally, MDL has intellectual roots in the algorithmic or descriptive complexity theory of Kolmogorov, Chaitin, and Solomonoff [73]. Only considering the probability distribution as a basis for generating descriptions, Rissanen endowed MDL with a rich information-theoretic interpretation [74]–[79]. Due to these characteristics, and it's reasonable to believe human brain meet minimum energy principle, causality analyzed with help of MDL principle may be more in line with physiological models of the brain. On the whole, our proposal takes the MDL principle as the single mathematical framework to select the generalized model for GCA, which to ensure the consistency, objectivity and the parsimony.

III. THE MINIMUM DESCRIPTION LENGTH PRINCIPLE

The principle of *parsimony* is the soul of model selection. To implement the *parsimony* principle, one must quantify *parsimony* of a model relative to the available data. With help of the work of Kolmogorov [80], [81], Wallace and Boulton [82], Rissenan formulated MDL as a broad principle

governing statistical modeling in general. At beginning of modeling, all we have is only the data. Luckily, With the help of the algorithmic or description complexity theory of Kolmogorov, Chaitin, and Solomonoff, MDL regards a probability distribution as a descriptive standpoint. And MDL has some connections with AIC and BIC, sometimes it behaves similarly to AIC and BIC [72], [83]. The difference is that MDL fixes attention on the length function rather than the code system. Therefore, many kinds of probability distribution can be compared in terms of their descriptive power [84]. Our code length guided framework can be used in generalized model selection as long as there is a probability distribution in the model.

A. PROBABILITY AND IDEALIZED CODE LENGTH
In order to describe the MDL principle explicitly, we deduced the formula of MDL in different cases. In model selection process of MDL, we need to compare the code length obtained by its probability distribution, so it is essential to understand the relationship between probability and the code length [72], [85].

A code ϱ on a set $\mathcal{A}$ is simply a mapping from $\mathcal{A}$ to a set of codewords. Let $\mathcal{A}$ be a finite binary set and let Q denote a probability distribution of the any element a in $\mathcal{A}$. The code length of a is that $-\log_2 Q$, the negative logarithm of Q. For example, the Huffman code is one of the algorithms that constructed this relationship between probability and idealized code length [84]. Suppose that elements of $\mathcal{A}$ are generated according a known distribution P. Given a code ϱ on $\mathcal{A}$ with length function L, the expected code length of ϱ with respect to P is defined as

$$L_\varrho = \sum_{x \subset A} P(x)L(x). \tag{5}$$

As is well known, if ϱ is prefix code, the code length L is equivalent to $-\log_2 Q(x)$ for some distribution Q on $\mathcal{A}$. There is given an arbitrary code, if no codeword is the prefix of any other, the unique decodability is guaranteed. Any code satisfying this codeword condition is called a prefix code.

By Shannon's Source Coding Theorem, for any prefix code ϱ on $\mathcal{A}$ with length function L, the expected L_ϱ is bounded below by $H(P)$, the entropy of P. That is

$$L_\varrho \geq H(P) = -\sum_{x \subset A} P(x)\log_2 P(x), \tag{6}$$

where equality holds if and only if $L = -\log_2 P$, in other words, the expected code length is minimized when $Q = P$.

B. CRUDE TWO-PART CODE MDL
In our view, modeling is a process that find the regularity of data and compress it. In model selection within MDL principle, What we have to do is selecting a suitable model based on the probability distribution of the object. Generally, the model we picked is overfitting or too simple. But model selection guided by the MDL principle, the complexity term or the error term in data fitting is incorporated into code

length guided framework, which ensures the objectivity of the operation.

Until now, there are several forms of MDL principle to polynomial or other types of hypothesis and model selection. But at the original MDL, it usually divides the modeling for the data set into two parts, one part is to describe the model's self-information. The other is to describe the data set with the help of chosen probability model in part one. Consequently, here we firstly introduce the most common implementation of the idea – the two part code version of MDL [72], [85].

Suppose the data $D \in \mathcal{X}^n$ where $\mathcal{X} = \{0, 1\}$. Then there is a probability $P \in \mathcal{M}$, and minimize

$$L_{1,2}(P, D) = L_1(D|P) + L_2(P). \tag{7}$$

Here, it will select a reasonable model for D to make good predictions of future data coming from the same source, which therefore models the data using the class $\mathcal{B}$ of all Markov chains [85].

1) THE FIRST PART
To get a better feel for the code L_1, we prepare to consider two examples. First, let $P_\theta \in \mathcal{B}^{(1)}$ be some Bernoulli distribution with $P_\theta(x = 1) = \theta$, and let $D = (x_1, \cdots, x_N)$. Since $P_\theta(D) = \prod P_\theta(x_i)$ and $\hat{\theta}$ is equal to the frequency of 1 in D, the first part $L_1(D|P)$ is given as

$$\begin{aligned} -\log P_\theta(D) &= -n_1 \log\theta - n_0 \log(1 - \theta) \\ &= -N[\hat{\theta}\log\theta + (1 - \hat{\theta})\log(1 - \theta)], \end{aligned} \tag{8}$$

where n_j denotes the number of occurrences of symbol j in D. Let $k = 2^\gamma$, the γth-order Markov chain model is denoted by $\mathcal{B}^{(k)}$, it's defined as

$$\mathcal{B}^{(k)} = \{P_\theta | \theta \in \Theta^{(k)}\}; \Theta^{(k)} = [0, 1]^k.$$

where $\theta = (\eta[1|0\ldots0], \eta[1|0\ldots01], \ldots, \eta[1|0\ldots11])$, for all n, x^n

$$P_\theta(D) = (\frac{1}{2})^\gamma \prod_{i=\gamma+1}^N P_\theta(x_i|x_{i-1}, \cdots, x_{i-\gamma}) \tag{9}$$

and

$$\begin{aligned} -\log P_\theta(D) = -N \sum_{y \in \{0,1\}^{\log k}} (&\hat{\eta}_{[1|y]}\log\eta[1|y] \\ &+ (1 - \hat{\eta}_{[1|y]})\log(1 - \eta_{[1|y]})) + \gamma. \end{aligned} \tag{10}$$

Here $\gamma = \log k$ is the number of bits needed to encode the first γ outcomes in D. The maximum likelihood parameters $\hat{\eta}_{[1|y]}$ are equal to the conditional frequencies of 1 prefixed by y.

2) THE SECOND PART
In order to describe a $P \in \mathcal{B}$, we have to describe a pair (k, θ). We encode the all parameter in the distribution model by firstly encoding k, which will use some prefix code C_{2a}, and then code θ with the help of the prefix code C_{2a}. The resulting code C_2(the code length of the first part) is then

defined by $C_2(k, \theta) = C_{2a}(k)C_{2b}(\theta|k)$. Firstly, N is nature number set,

$$L_{2a}(k) = L_N(k) = O(\log k). \quad (11)$$

Since $\Theta^{(k)} = [0, 1]^k$ is an uncountable infinite set, we would have to discretize it firstly. More precisely, we will restrict $\Theta^{(k)}$ to some finite set $\ddot{\Theta}_d^{(k)}$ of parameters that can be described using a finite precision of d bits per parameter.

Now that the number of elements of $\ddot{\Theta}_d^{(k)}$ is $(2^d)^k$. Thus it needs $\log(2^d)^k = k \cdot d$ bits to describe any particular $\vartheta \in \ddot{\Theta}_d^{(k)}$, and may call d the precision used to encode a parameter. Letting $d\vartheta$ be the smallest d such that $\theta \in \ddot{\Theta}_d^{(k)}$, this gives

$$L_2(k, \theta) = \begin{cases} L_N(k) + L_N(d_\vartheta) + kd_\vartheta, & \text{if } d_\vartheta < \infty \\ \infty, & \text{otherwise.} \end{cases} \quad (12)$$

At the end, the crude two-part code MDL for Markov chain hypothesis selection is given by,

$$\min_{k, d \in N; \theta \in \ddot{\Theta}_d^{(k)}} -\log P_{k,d}(D) + kd + L_N(k) + L_N(d). \quad (13)$$

IV. CODE LENGTH GUIDED MODEL SELECTION IN CAUSALITY ANALYSIS

As stated above, conventional GCA splits the whole process into two stages which are actually modeling endogenous information and exogenous information. Since MDL has a close relationship with information theory, causality analysis with MDL here will also be more convincing and suitable. Further the regression of endogenous information can be converted to code length, and relative effect of exogenous variables can be also quantified by the code length guided framework, which the whole model selection process for GCA is unified into same model space.

The following is that MDL principle guided model selection in causality analysis [68], and variable X_N is given,

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_n x_{t-n} + \epsilon_t. \quad (14)$$

where $t = 1, \cdots, m$, which m can be anyone more than n to keep the solution determined, N is the data length. The parameter vector consists of data $\theta = (n, \xi)$ and $\xi = (\sigma^2, a_1, \cdots, a_n)$, where $\sigma^2 = \xi_0$ is the variance-parameter of zero-mean Gaussian distribution model for ϵ_t. In order to describe x_t, turn to Gaussian distribution for ϵ_t. Clearly, to describe the Gaussian distribution model, applying the two part form of MDL, the total code length is given as

$$L(x, \theta) = m \ln \sqrt{2\pi}\sigma + \frac{RSS}{2\sigma^2} + \sum_{i=0}^{n} \ln \frac{|\xi_i|}{\delta} + \ln(n + 1). \quad (15)$$

where $RSS = \sum_{t=1}^{m}(\epsilon_t^2)$ denotes the residual sum of squares corresponding to the estimation in model. And δ is the precision, it's optimal to choose $1/\sqrt{N}$ [72], [75], [86]. Specially, $\frac{|\xi_i|}{\delta} < 1$ should be ignored.

A. TIME-DOMAIN FORMULATION

Combining with the above formula, the causality investigation with the code length guided framework in the time domain can be carried out. There are two time-series X_N and Y_N, then we consider two different models A and B (in Eq. (16) and Eq. (17)) to describe x_t. The representations are

$$\begin{cases} X_t = \sum_{j-1}^{n} a_{1i} X_{t-j} + \epsilon_{1t} \\ Y_t = \sum_{j-1}^{n} d_{1i} Y_{t-j} + \eta_{1t} \end{cases} \quad (16)$$

where $var(\epsilon_{1t}) = \Sigma_1$ and $var(\eta_{1t}) = \Gamma_1$. Bivariate regressive representations are given,

$$\begin{cases} X_t = \sum_{j-1}^{n} a_{2i} X_{t-j} + \sum_{j-1}^{n} b_{2i} Y_{t-j} + \epsilon_{2t} \\ Y_t = \sum_{j-1}^{n} c_{2i} X_{t-j} + \sum_{j-1}^{n} d_{2i} Y_{t-j} + \eta_{2t} \end{cases} \quad (17)$$

where $var(\epsilon_{2t}) = \Sigma_2$ and $var(\eta_{2t}) = \Gamma_2$, and their contemporaneous covariance matrix is

$$\Sigma = \begin{pmatrix} \Sigma_2 & \Upsilon_2 \\ \Upsilon_2 & \Gamma_2 \end{pmatrix} \quad \text{where } \Upsilon_2 = cov(\epsilon_{2t}, \eta_{2t}).$$

Finally, ϵ_{1t} and ϵ_{2t} have Gaussian distribution with mean 0 and unknown variance σ^2, which denote the *noise* of time-series are fitting residual. Therefore, the distribution of residual terms ϵ_t can be a standpoint to describe the model within MDL. Then, the code length of model A and B we obtained can be compared to identify the causal influence between x_t and y_t. In the whole process of model selection for GCA, the causality investigation was mapped into the unified code length guided framework. According to Eq.(15), the code length of model A and model B can be given respectively. By the definition of Granger causality, the influence from Y to X is defined by our code length guided framework,

$$F_{Y \to X} = L_X - L_{X+Y}. \quad (18)$$

where L_X denotes the code length of optimal model in Eq.(16), and L_{X+Y} denotes the code length of optimal model in Eq.(17). If $F_{Y \to X} > 0$, it means causal influence from Y to X existed. Otherwise, there is no causal influence existed from Y to X. As the causality represented above, our proposal can unify two-stage scheme into the code length guided framework, which can avoid inconsistency of two different mathematical theories or the subjectivity of F-statistics in model selection of the conventional GCA.

To compare conditional GCA, we consider the influence from Y to X while controlling for the effect from conditional node Z to X. Firstly, the joint autoregressive representation is given

$$\begin{cases} X_t = \sum_{j-1}^{n} a_{3i} X_{t-j} + \sum_{j-1}^{n} b_{3i} Z_{t-j} + \epsilon_{3t} \\ X_t = \sum_{j-1}^{n} a_{4i} X_{t-j} + \sum_{j-1}^{n} b_{4i} Y_{t-j} + \sum_{j-1}^{n} c_{4i} Z_{t-j} + \epsilon_{4t} \end{cases} \quad (19)$$

and $var(\epsilon_{3t}) = \Sigma_3$, $var(\epsilon_{4t}) = \Sigma_3$. By the definition of conditional GCA, if $F_{Y \to X} > 0$ existed, causal influence from Y to X conditioned Z is defined

$$F_{Y \to X|Z} = L_{X+Z} - L_{X+Y+Z}. \quad (20)$$

Same as above, if $F_{Z \to X} > 0$ existed, causal influence from Z to X conditioned Y is given

$$F_{Z \to X | Y} = L_{X+Y} - L_{X+Y+Z} \qquad (21)$$

Clearly, in our code length guided framework, all candidate models can be compared freely in the context of their code length. Different from the traditional method, the causal connection is obtained by repeated pairwise comparison between models, our method can map all candidate models into the same model space without the repeating comparison and to obtain the conditional causal influence directly. Which is, if both $F_{Y \to X} > 0$ and $F_{Z \to X} > 0$ existed,

$$F_{Y,Z \to X} = min(L_{X+Y}, L_{X+Z}) - L_{X+Y+Z}. \qquad (22)$$

Here if $F_{Y,Z \to X} > 0$ existed, it means that both Y and Z have direct influence on X. But if $F_{Y,Z \to X}$ is less than 0, there will be two cases. One is $F_{Y,Z \to X} = (L_{X+Y} - L_{X+Y+Z}) < 0$ existed, it means only Y has direct influence on X. The other is $F_{Y,Z \to X} = (L_{X+Z} - L_{X+Y+Z}) < 0$ existed, it means that Z impacts X directly. In the unified model space, multiple selected models can be directly compared by code length, which can release the complexity of the algorithm. In this way, our proposal is more in line with Occam's razor, or the principle of *parsimony*.

B. FREQUENCY-DOMAIN FORMULATION

With help of Geweke's work [39], the total interdependence between two time series X_t and Y_t can be decomposed into three components: two directional causal influences due to their interaction patterns, and the instantaneous influence due to factors possibly exogenous to the (X, Y) system (e.g. a common driving input) [41], [54]. Here other forms of regressive representations need to be considered, We first rewrite Eq.(16) and Eq.(17)

$$\begin{pmatrix} a_2(L) & b_2(L) \\ c_2(L) & d_2(L) \end{pmatrix} \begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} \epsilon_{2t} \\ \eta_{2t} \end{pmatrix} \qquad (23)$$

where $a_2(0) = 1, b_2(0) = 0, c_2(0) = 0, d_2(0) = 1$, the lag operator L denotes $LX_t = X_{t-1}$. Performing Fourier transform on both sides of Eq.(23), then we left-multiply

$$P = \begin{pmatrix} 1 & 0 \\ -\dfrac{\Upsilon_2}{\Sigma_2} & 1 \end{pmatrix} \qquad (24)$$

on both sides and rewrite the result equation, the normalized equations yield

$$\begin{pmatrix} X_\omega \\ Y_\omega \end{pmatrix} = \begin{pmatrix} D_{11}(\omega) & D_{12}(\omega) \\ D_{21}(\omega) & D_{22}(\omega) \end{pmatrix} \begin{pmatrix} E_2(\omega) \\ H'_2(\omega) \end{pmatrix}, \qquad (25)$$

where $H'_2(\omega) = H_2(\omega) - \frac{\Upsilon_2}{\Sigma_2} E_2(\omega)$. The spectral matrix is $S(\omega) = D(\omega)\Sigma D^*(\omega)$, where $*$ denotes complex conjugate and matrix transpose. The spectrum of X_t is

$$S_{11}(\omega) = D_{11}(\omega)\Sigma_2 D_{11}^*(\omega) - D_{12}(\omega)\Upsilon_2' D_{12}^*(\omega), \quad (26)$$

where $\Upsilon_2' = \Gamma_2 - \frac{\Upsilon_2}{\Sigma_2}\Upsilon_2$. The first term in Eq.(26) is represented as the intrinsic influence and the second term as

the causal influence of X_t due to Y_t at frequency ω. Based on this transformation, the causal influence from Y_t to X_t at frequency ω is

$$f_{Y \to X}(\omega) = ln \frac{S_{xx}(\omega)}{D_{11}(\omega)\Sigma_2 D_{11}^*(\omega)}. \qquad (27)$$

The model orders of historical information in frequency domain for conventional GCA, are determined by AIC/BIC. Distinct from conventional GCA, we obtained the causal connectivities between nodes at frequency ω by the code length guided framework.

V. EXPERIMENTS
A. 3-NODE NETWORK SIMULATION EXPERIMENTS
1) PROTOCOL FOR 3-NODE NETWORK
To verify the performance of MDL principle in synthetic data experiment, a simple 3-node network is enabled. There were 1000 data points in time series of each node, to keep the stationarity of synthetic data, the first 700 data points were removed. The initial value of each node was 1, the variance of noise term $\epsilon_i (i = 1, 2, 3)$ varied from 0.15 to 0.35. The nodes was generated by

$$\begin{cases} y_{1,t} = 1.5y_{1,t-1} - 0.9y_{1,t-2} + \epsilon_1 \\ y_{2,t} = 0.8y_{1,t-1} + 0.2y_{2,t-1} + \epsilon_2 \\ y_{3,t} = -0.8y_{1,t-1} + 0.4y_{3,t-1} + \epsilon_3 \end{cases} \qquad (28)$$

2) RESULTS FROM 3-NODE NETWORK
Firstly, to ensure the rigor of the proposal, the distribution of residual between the selected model and the observed data was verified. It was almost corresponding Gaussian distribution with mean 0, seen in Fig. 2(a) and 2(b), which guaranteed the validity for our initial description standpoint. And code length changed with different time lag in autoregressive model had shown in Fig. 2(c), it dropped down to the minimum when time lag was 2, which corresponded with generated model in Eq.(28).

Then, the causal connectivities identified by our proposal had shown in Fig. 3(b), which obtained causal connection by comparing the code lengths according to Eq. (15), then the causal connectivities obtained by conventional GCA had shown in Fig. 3(a). We found that causal influences from node 1 to node 2 and node 3 were identified both in GCA and our proposal. In other four causal influences, our proposal almost guaranteed 99% accuracy. But conventional GCA only guaranteed 95% accuracy.

At the same time, experiments with three different noise levels and three confidence intervals had been carried out, the comparison between conventional GCA and our proposal showed in Table. 1. The accuracy of causal connection associated with each node and the overall true model was also verified. Comparing the accuracy of the single node and the overall network at the same noise level and the same confidence interval, it's unevenly distributed in conventional GCA. Meanwhile, the conventional GCA were very sensitive due
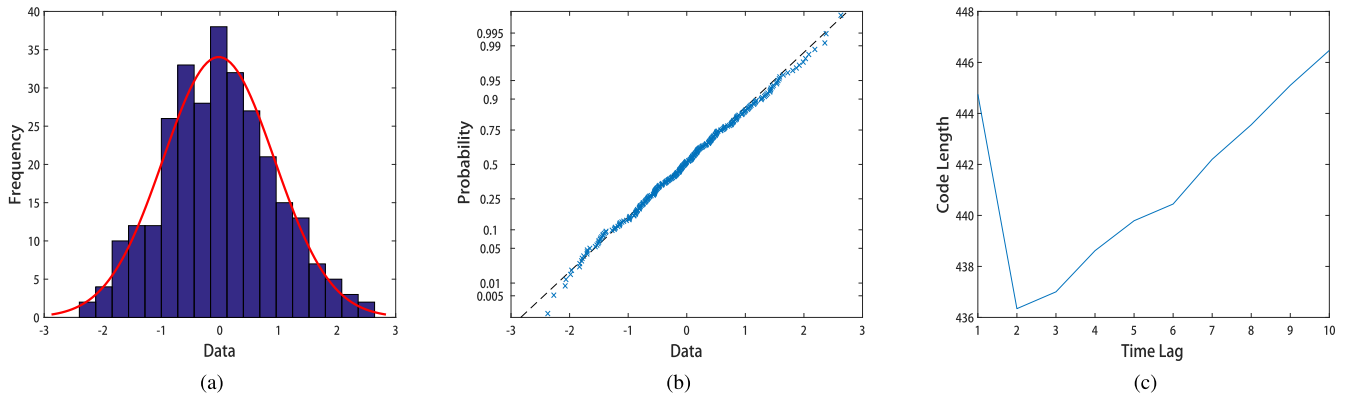
FIGURE 2. The fitting Guassian distribution of the residual. (a): a histogram of values in data and fits a normal density function. (b): comparing the distribution of the data to the normal distribution. (c): code length obtained by our proposal, changed with different time lag in node 1.

TABLE 1. Comparison between conventional GCA and MDL method in 3-node network.

Noise level	Method		Accuracy(%)			
			Node 1	Node 2	Node 3	Total
Low	F-statistics	$\alpha = 0.1$	82.3	71.8	72.1	65.4
		$\alpha = 0.05$	90.9	85.5	85.5	82
		$\alpha = 0.01$	98.1	96	96.2	95.2
	MDL		98.4	97.3	97.2	96.5
Moderate	F-statistics	$\alpha = 0.1$	81.4	72.5	71.8	64.8
		$\alpha = 0.05$	89.4	84.6	84.7	79.9
		$\alpha = 0.01$	97.6	96.2	96.3	95.1
	MDL		98.5	97.7	97.4	96.8
High	F-statistics	$\alpha = 0.1$	80.3	72	71.7	64.5
		$\alpha = 0.05$	89.5	83.6	83.7	79.1
		$\alpha = 0.01$	98	96.7	96.7	95.7
	MDL		98.8	98	97.6	97.2

The variance of the low noise level data ranges from 1.5 to 3.5, and the moderate (high) level data ranges from 2.5 to 4.5 (3.5-5.5). $1 - \alpha$ denotes the significance level of F-test in conventional GCA. The accuracy of node i (i=1,2,3) only measures the causal connectivities with node i, not including the connections between the other two nodes. And the total accuracy represents the accuracy that the true model is found, it means only quantifies the causal influence from node 1 to node 2 and node 3.
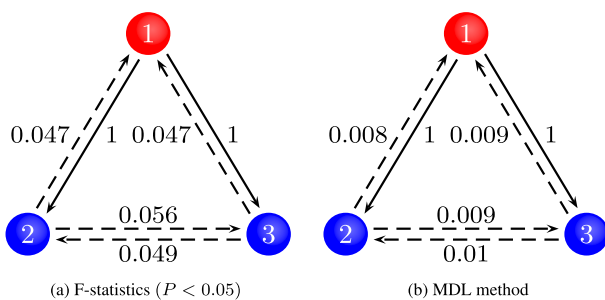

(a) F-statistics ($P < 0.05$) (b) MDL method

FIGURE 3. Comparison between two methods within low variance noise (0.15-0.35), the numbers on the arrow indicated the accuracy that the causal influence was identified in 1000 samples. Specially, the accuracy was only represent the probability that the single connection direction was identified by two methods. (a): causal network obtained by conventional GCA. (b): causal network obtained by our proposal.

to the different confidence intervals, especially in the identification of the overall network, even it had a stable performance at different noise level. In generally, our proposal always showed a relatively good performance whether it's

identification in the whole network or a single node. It was worth noting that the results obtained by conventional GCA when $\alpha = 0.01$ in *F*-statistics were close to the results of our proposal. The results were also in line with our expectations, that was because our proposal consider the complexity of the model more thoughtfully. As we emphasized above, there is only one mathematical principle guiding the model selection throughout the GCA process, thus our proposal is a more rigorous approach or a more robust approach.

The oscillations in neurophysiological systems and neuroscience data are thought to constrain and organize neural activity within and between functional networks across a wide range of temporal and spatial scales. Geweke-Granger causality demonstrated that the oscillations at specific frequencies had been associated with the activation or inactivation of different encephalic region. But for conventional GCA in frequency domain, model selection of history information is determined by AIC or BIC. For our code length guided framework, the selected models for history information will regress into the *true* model space automatically. Due to its
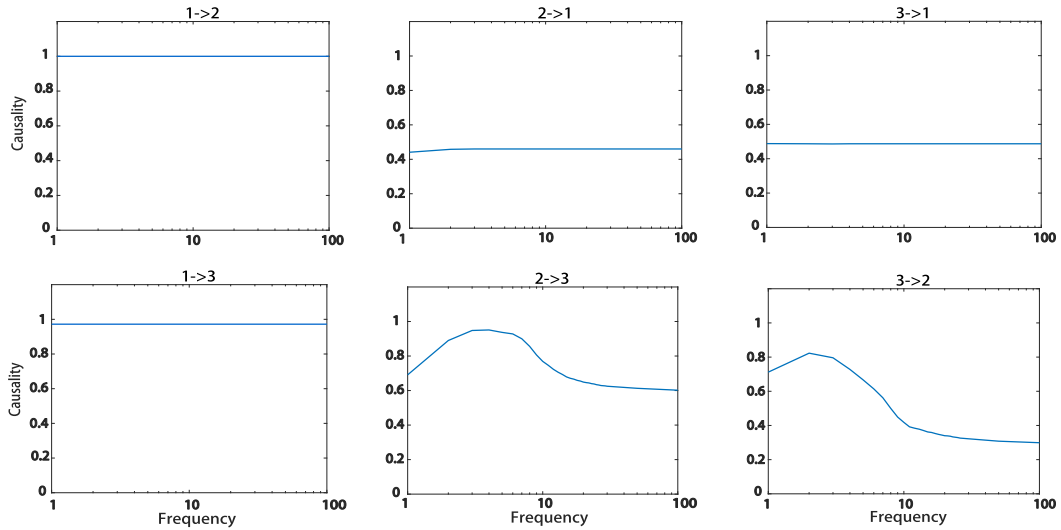
FIGURE 4. Causal connectivities between 3-node network obtained in frequency domain. The causal influence was obtained at frequency ω range from $1 \sim 30$, and at 50,100Hz.
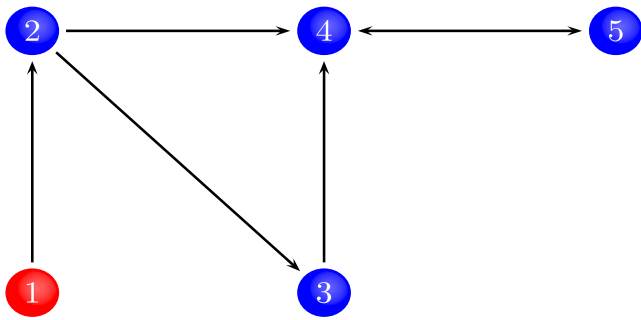

FIGURE 5. The relationships of simulation data sets in 5-node network.

intellectual roots in descriptive complexity and close tie with information theory, our proposal may be more capable to identified causality in frequency domain.

Same as time domain, causal connectivities at frequency ω between 3-node network obtained by our proposal showed in Fig. 4. In particular, causal connectivities in frequency domain were obtained within two nodes, which meant that we did not introduce conditional GCA in the frequency domain. This is mainly because there seems to be still some obfuscation with the method of using conditional GCA concept in the frequency domain. Therefore, as shown in Fig. 4, some non-existent connections between nodes were often misjudged, except for the causal influence $1 \rightarrow 2$ and $1 \rightarrow 3$. But we found that causal connectivities between node 2 and node 3 had a bigger chance to be misjudged at low frequency $(0 - 10$ Hz). Actually, comparing results obtained by conventional GCA in the time domain, causalities between two nodes were more legible in frequency domain, which meant only direct causalities showed more consistent results in our analysis. For example, there were only stable and significant causal influence existed in $1 \rightarrow 2$ and $1 \rightarrow 3$.

B. 5-NODE NETWORK SIMULATION EXPERIMENTS

1) PROTOCOL FOR 5-NODE NETWORK

To verified in more detail whether the proposed MDL method is just a conventional GCA method with a higher level of confidence, a complex 5-node network was given to further verified the robustness and the validity of our proposal, seen in Fig. 5. Noise terms $\epsilon_i (i = 1, 2, \ldots, 5)$ were the Guassian distribution with mean 0, and the variance ranged from 0.15 to 0.3. The first two initial values of nodes are 1, they were given by

$$
\begin{cases}
x_{1,t} &= 0.792x_{1,t-1} - 0.278x_{1,t-2} + \epsilon_1 \\
x_{2,t} &= 0.768x_{2,t-1} - 0.503x_{2,t-2} + 0.83x_{1,t-1} \\
&\quad -0.32x_{1,t-2} + \epsilon_2 \\
x_{3,t} &= 0.67x_{3,t-1} - 0.312x_{3,t-2} + 0.56x_{2,t-1} \\
&\quad -0.42x_{2,t-2} + \epsilon_3 \\
x_{4,t} &= 0.733x_{4,t-1} - 0.27x_{4,t-2} + 0.72x_{2,t-1} \\
&\quad -0.27x_{2,t-2} + 0.52x_{3,t-1} - 0,456x_{3,t-2} \\
&\quad +0.76x_{5,t-1} - 0.33x_{5,t-2} + \epsilon_4 \\
x_{5,t} &= 0.845x_{5,t-1} - 0.24x_{5,t-2} + 0.68x_{4,t-1} \\
&\quad -0.254x_{4,t-2} + \epsilon_5
\end{cases}
\tag{29}
$$

2) RESULTS FROM 5-NODE NETWORK

And in other connections in 5-node network, the accuracy of our proposal also was not below 98.4%. Whereas, conventional GCA did not showed the same robustness of our proposal. Causal connectivities between direct related nodes was not well identified, for example only 507 samples were identified as causal influences from node 3 to node 4 in 1000 samples and causal influence from node 5 to node 3 was identified at 92.9% accuracy. And in other connections, the accuracy of conventional GCA was more poor than our proposal. Clearly, in higher confidence levels ($\alpha = 0.01$) of conventional GCA,
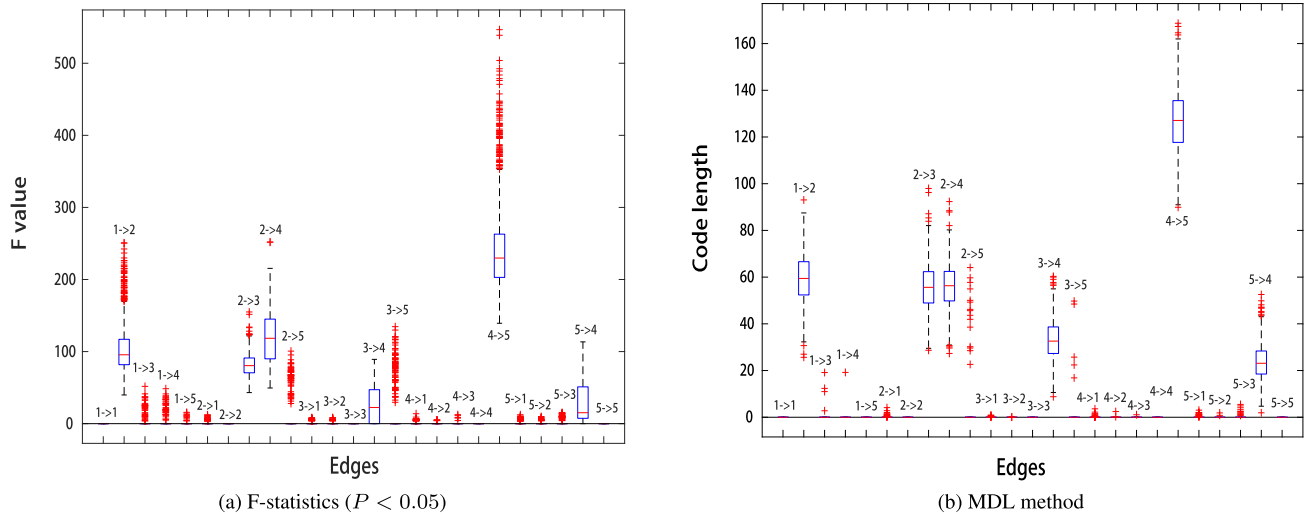
(a) F-statistics ($P < 0.05$)

(b) MDL method

FIGURE 6. Causal connectivities between 5 nodes identified by conventional GCA and our proposal respectively. (a): causal connectivities was obtained by conventional GCA at $\alpha = 0.05$ in F-test. (b): causal connectivities obtained by our proposal.

TABLE 2. 5-Node networks.

	Edge	F-statistics($\alpha = 0.05$)	($\alpha = 0.01$)	MDL method
true connections	$1 \longrightarrow 2$	1000/1000	(1000/1000)	1000/1000
	$2 \longrightarrow 3$	1000/1000	(1000/1000)	1000/1000
	$2 \longrightarrow 4$	1000/1000	(1000/1000)	1000/1000
	$3 \longrightarrow 4$	507/1000	(456/1000)	1000/1000
	$4 \longrightarrow 5$	1000/1000	(1000/1000)	1000/1000
	$5 \longrightarrow 4$	929/1000	(780/1000)	1000/1000
false connections	$1 \longrightarrow 3$	36/1000	(8/1000)	4/1000
	$1 \longrightarrow 4$	43/1000	(1/1000)	1/1000
	$1 \longrightarrow 5$	11/1000	(0/1000)	0/1000
	$2 \longrightarrow 1$	54/1000	(12/1000)	15/1000
	$2 \longrightarrow 5$	58/1000	(14/1000)	15/1000
	$3 \longrightarrow 1$	44/1000	(17/1000)	7/1000
	$3 \longrightarrow 2$	15/1000	(2/1000)	2/1000
	$3 \longrightarrow 5$	83/1000	(29/1000)	5/1000
	$4 \longrightarrow 1$	30/1000	(8/1000)	14/1000
	$4 \longrightarrow 2$	6/1000	(0/1000)	2/1000
	$4 \longrightarrow 3$	5/1000	(1/1000)	1/1000
	$5 \longrightarrow 1$	53/1000	(18/1000)	16/1000
	$5 \longrightarrow 2$	17/1000	(1/1000)	2/1000
	$5 \longrightarrow 3$	41/1000	(8/1000)	11/1000

Causal connectivities between nodes in 5-node network within low variance noise (0.15-0.3). True connection represents the causal influence truly existed in Fig. 5, false connection means not existed.

although the specificity of causal identification increased, its sensitivity decreased.

The result of connections between 5 nodes showed in Fig. 6(b), causal influence analyzed by code length was largely consistent with connections in Fig. 5. Simultaneously, causalities analyzed by conventional GCA between 5 nodes showed in Fig. 6. Same as results in 3-node network, our proposal showed 100% consistency with connections in Fig. 5 between directly causal related nodes, seen in Table. 2.

More importantly, whether the causal influence from Node 3 to Node 4 or from Node 5 to Node 4, the significance level of connection were not enough to be identified, even for $\alpha = 0.01$ in F-statistics. Therefore, the results demonstrated that our proposal was not equivalent to the conventional GCA with a higher significance level in F-statistics at all.

Obviously, when target network was more complicated in simulation, our proposal showed a more desirable property in time domain, while conventional GCA generally made mistakes. Luckily, our proposal performed very well regardless of the existence fo relationships between nodes, even as in more complicated network. Same as 3-node network, the result causal network was unrelated with the varying variance of noise from (0.15-0.3) to (0.35-0.5). In conditional causality analysis, our approach reduced the complexity of algorithm, eliminating the need for repeated pairwise comparisons between models, while ensured the accuracy of results.

Subsequently, we identified the causal connectivities among 5-node network in frequency domain by our proposal, seen in Fig. 7. Same as the 3-node network, causal
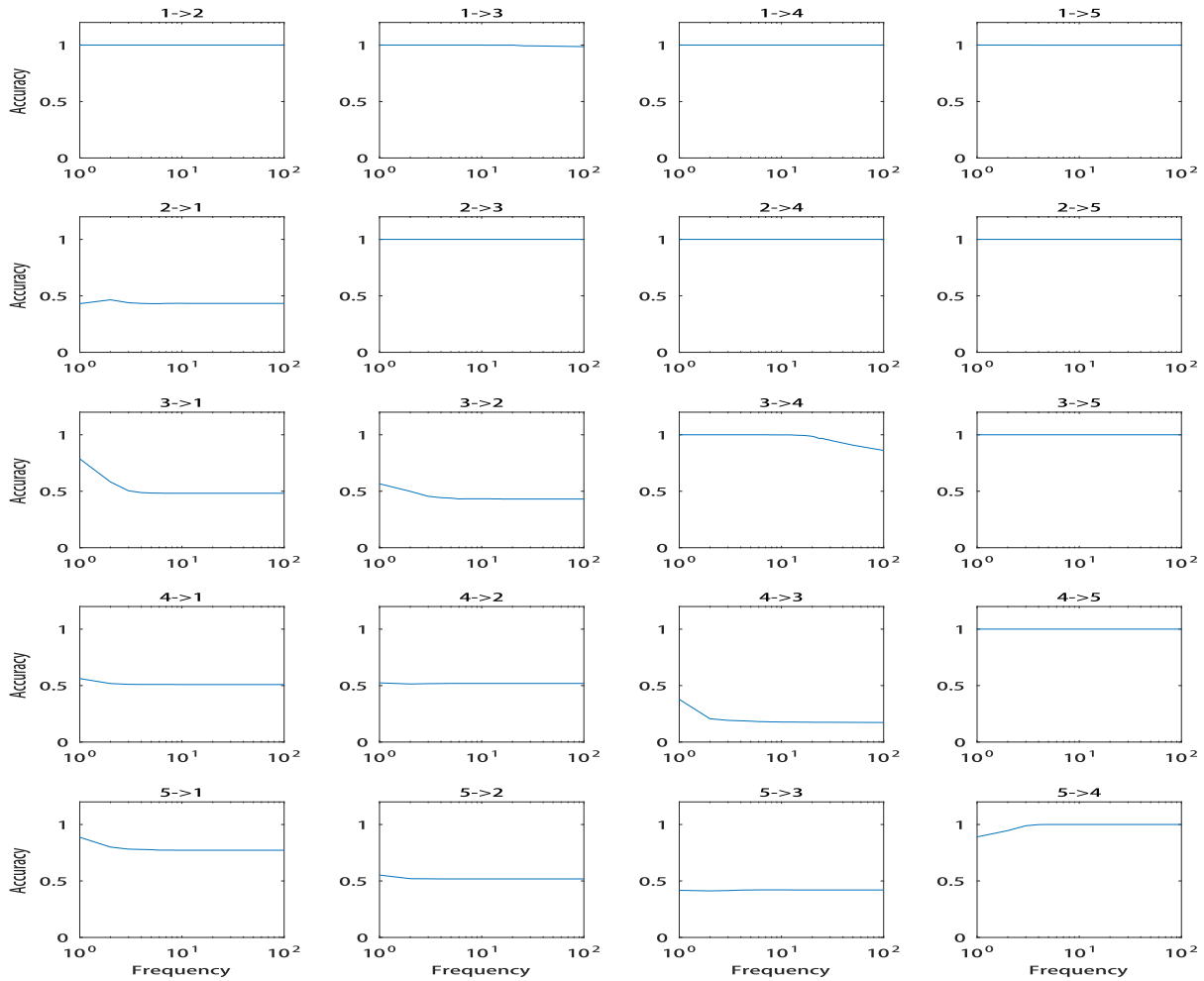
FIGURE 7. Causal connectivities of 5-node network obtained in frequency domain. Same as 3-node network, the causal influence was identified at frequency ω range from $1\sim 30$, and at 50, 100Hz.

connection networks were identified without introducing conditional GCA, and the causal connection network had regular characteristics. The causal influence whether it was direct or indirect existed in 5-node network was more stable to be identified in the frequency domain. Similarly, since conditional GCA was not considered, other non-existent causal connectivities had a chance to be identified. And in 5-node network, the possibility of being misjudged was even greater. Therefore, it is necessary to introduce conditional GCA to distinguish direct from indirect influences between system components in the frequency domain. And at the same time, we found that removing the noise frequency component is an obstacle to causality analysis, main reason is that we have no prior knowledge about which one is *noise* or others in row data.

C. REAL DATA
1) EXPERIMENTAL PROTOCOL
In the study we let ten subjects perform simple one-digit(consisting of 1-10) serial addition (SSA) and complex two-digit(consisting of 1-5) serial addition (CSA) by visual

stimulus and simultaneously measured their brain activities with fMRI.

Immediately, we asked the subjects to perform same serial addition arithmetic tasks by auditory stimulus. Nine right-handed healthy subjects (four female, 24±1.5 years old) and one left-handed healthy female subject (24 years old) participated. One of the subject's(a right hand male) data was deleted due to excessive head motion. All subjects volunteered to participate in this study with informal written consent by themselves.

The causal connection network of one subject performing same task should be same or similar, which is also the basic assumption in group analysis. Therefore, for mental calculation under different stimuli, causal connection network of same subject may be different in the input node of stimulus, but we have reason to postulate that the causal connection should be same inside the mental calculation network, at least it should be similar. Thus, we will compare the similarity of causal connection network within the same mental calculation task under different stimuli. More directly, in order to quantify which method is more robust, we will compare the
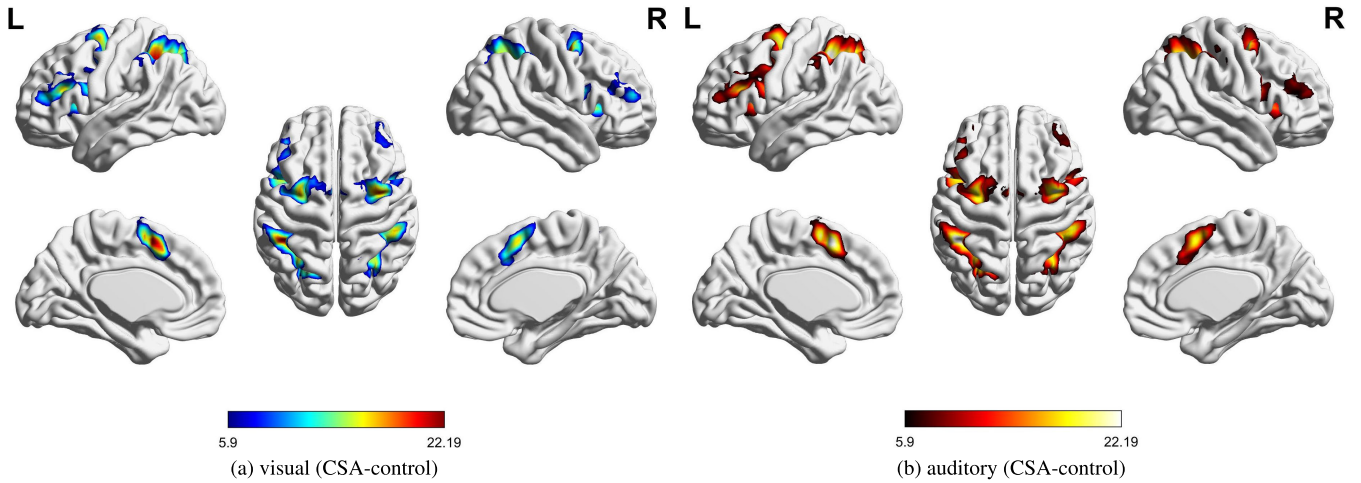
(a) visual (CSA-control)

(b) auditory (CSA-control)

FIGURE 8. Mental calculation of CSA-control state under the two stimuli (visual stimulus and auditory stimulus), the activation regions were processed by SPM12, the control state meant that the sample was in rest state without mental calculation. (a): CSA-control state under visual stimulus. (b): CSA-control state under auditory stimulus.(P<0.0001, uncorrected).
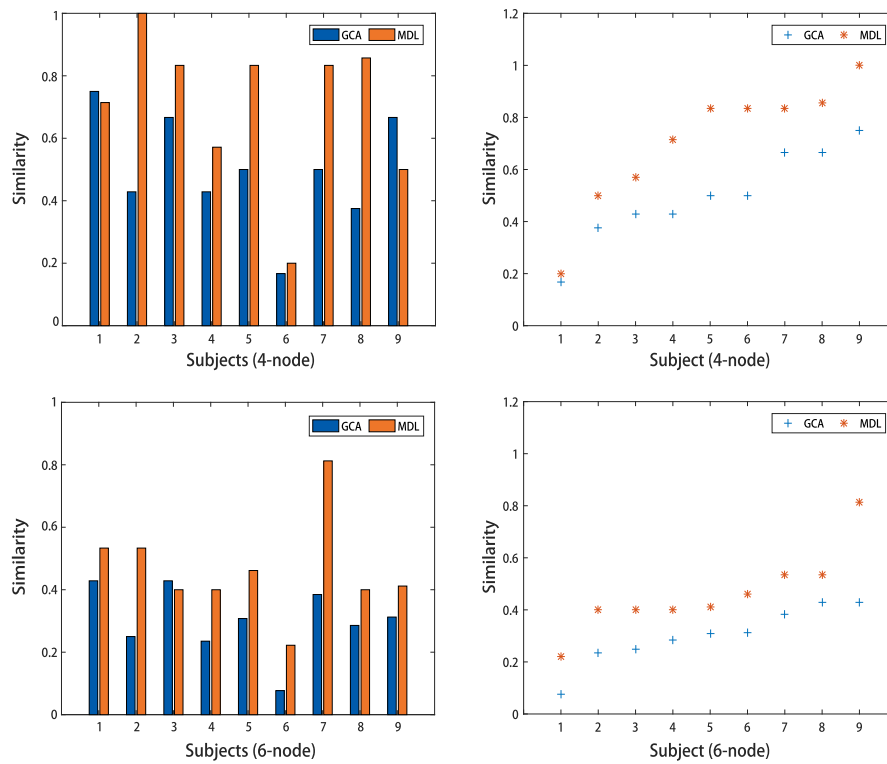

FIGURE 9. The similarity in 4-node mental calculation network and 6-node full network respectively. Left panel compared the similarity on the individual, and the right panel compared the distribution of similarities, collectively.

similarity of the result network within auditory and visual stimulus on each subject. To obtain the similarity between networks, we define a measure method to quantify the similarity,

$$S = \frac{\sum \sum (A \cap B)}{\sum \sum (A \cup B)}. \tag{30}$$

where S represents the measured similarity, A and B are the connected matrix of mental calculation networks. The numerator is the sum of intersections of the same connected edges, and the denominator is the sum of the unions of the connected edges.
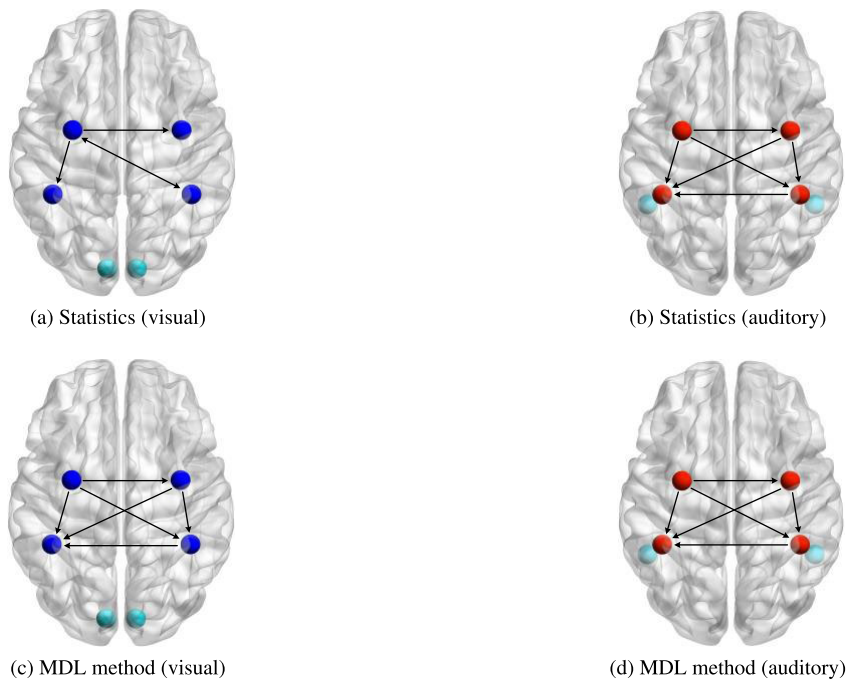
(a) Statistics (visual)

(b) Statistics (auditory)

(c) MDL method (visual)

(d) MDL method (auditory)

FIGURE 10. Causal connectivities of subject 2 in mental calculation network under two stimuli. Cyan node represented visual/auditory stimulus node. The blue/red nodes related to the mental calculation network stated in the text. (a) and (b): mental calculation network obtained by conventional GCA. (c) and (d): mental calculation network obtained by proposed approach.

(a) Statistics (visual)

(b) Statistics (auditory)

(c) MDL method (visual)

(d) MDL method (auditory)

FIGURE 11. Causal connectivities of subject 7 in mental calculation network under two stimuli. Cyan node represented visual/auditory stimulus node. The blue/red nodes related to the mental calculation network stated in the text. (a) and (b): mental calculation network obtained by conventional GCA. (c) and (d): mental calculation network obtained by proposed approach.
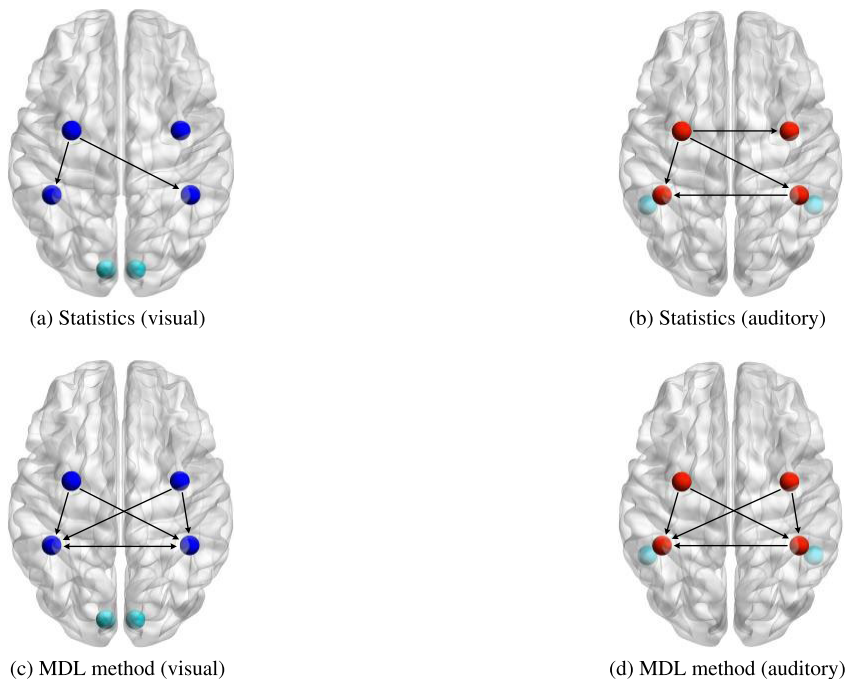
2) MENTAL CALCULATION NETWORK

We already have verified the validity of MDL in simulation data, but behavior in real data will determine the truly

robustness of the methods. Firstly, the activation regions of mental calculation under different stimuli showed in Fig. 8, and as had been postulated above, the activation regions

between two senses were identical. The similarities in 4-node connection network and 6-node full connection network showed in Fig. 9 respectively. In 4-node network, we found that causal connection networks of mental calculation between visual stimulus and auditory stimulus were very similar for most subjects, except for subject 6. There were seven of nine subjects the similarities were above 0.6. Turned to conventional GCA, we found that only three subjects of causal connection networks between the two stimuli had a similarity above 0.6. Meanwhile, only in subject 1 and subject 9, we found the similarity of causal networks between two senses was above our proposal. Even for these two subjects, the similarities in our proposal were close to conventional GCA, especially in subject 1. Further the similarities in 6-node full connection network also showed in Fig.11. Duo to the difference between the input node of stimulus, the similarities in two methods were not above 0.5 mostly, but there were still 3 subjects the similarities in our proposal were mostly above 0.5, especially for subject 7. Clearly, whether inside connection network of mental calculation or among the 6-node full connection network, the similarities in most of our subjects identified by our proposal were more than conventional GCA.

Then, causal mental calculation networks of two subjects showed in Fig. 10 and Fig. 11 respectively. As stated above, the input node of stimulus should not be included into the network to be compared, we removed the causal connection between stimuli nodes and other four nodes of mental calculation network. Obviously, our proposal also had a desirable robustness in fMRI data. As seen in Fig. 10 and Fig. 11, for subject 2 and subject 7, the mental calculation networks under different stimuli were almost identical. By the way, even for different subjects, the causal connection networks were similar at a large extent. As for conventional two-stage GCA, causal connection networks of two subjects above were more irregular, which seen in Fig. 10 and Fig. 11. Comparing the similarity between causal networks identified by our proposal, the networks obtained by conventional GCA had almost no consistency characteristics, the networks under different stimuli appeared to be unrelated. Conventional two-stage GCA identified a inconclusive results in mental calculation. Consistent with the results in simulation, our proposal was more robust in identifying causal connection network, especially in complex networks.

VI. CONCLUSION AND PERSPECTIVE

A. CONTRIBUTIONS AND DISCUSSIONS

The novelty of the present study is not in including the MDL principle in the GCA procedure, but rather in considering the MDL principle as unity strategy for model selection in the analysis procedure. Conventional GCA usually consists of two stages: (1) AIC/BIC for the predictors associated with internal or external information, and (2) F-statistic for evaluating the relative effects of exogenous variables. We emphasize that these two parts fall within the scope of

model selection, and distinct theory might generate different model selection criteria. Model selection should follow the same mathematical theory in the GCA process.

In this paper, we have addressed this concern and have proposed a unified model selection approach based on the MDL principle for GCA in the context of the general regression model paradigm. We have demonstrated its efficacy over conventional two-stage GCA approach in a 3-node network and a 5-node network synthetic experiments. All results confirm the superiority of proposed approach over conventional two-stage GCA. The unified model selection approach is capable of identifying the actual connectivity in all the cases, at the same time, avoiding greatly the false influences caused by the presence of noise. GCA was originally designed to handle pairs of variables, and may produce misleading results when the true relationship involves three or more variables [31], [32]. This case occurs when single connections are strong enough: if $A \rightarrow B$ and $B \rightarrow C$, then very likely $A \rightarrow C$ will be picked up by GCA, thus inducing a denser connected network than the truth. Our results suggest that the proposed approach can remove the indirect connection and retain the direct connection. The use of a better model selection strategy can improve the performance of GCA in systems involving a lager number of variables. As noted earlier (section II), for the statistical strategy, the resulting network are heavily dependent on the significance level chosen. In the simulation experiments involving a 5-node system, while a moderate significance level ($P < 0.05$) produces higher sensitivity in revealing actual connectivity (true positive, TP), it has lower specificity in avoiding false connectivity (true negative, TN). A more stringent significance threshold ($P < 0.01$) makes the reverse effect. In contrast, the proposed approach maintains high sensitivity and specificity in all cases, indicating that the MDL principle do much more than a higher significant level in F-statistics. The results illustrate the benefits of performing comparisons over all competing models in the model space as well.

More importantly, the proposed approach obtained more consistent results in a challenging fMRI dataset, in where visual/auditory stimuli with the same presentation design give identical neural correlates of mental calculation, allowing one to evaluate the performance of different GCA methods. This provides clear experimental proof that unified GCA is superior to conventional GCA [87]. The results are also consistent with the current consensus that model selection is crucial to investigate causal connectivity using neuroimaging techniques [88].

Essentially, both the MDL and statistical strategies, indeed all model selection methods, attempt to seek a trade-off between goodness-of-fit and complexity of the model involved. The MDL uses the code length which describes both the model complexity and the fitting error to achieve such a trade-off, whereas the statistical strategy implicitly conveys it in F-distribution function through degrees of freedom. In the MDL framework all competitive models can directly compare in terms of description length, without

the need for an intermediate model. In contrast, *F*-statistics resorts the extra sum of squares principle to determine which model gives a better fit to the data. This approach requires one model (the restricted model) to be nested within another model (the unrestricted model). We argue that the pairwise comparison in *F*-statistics is responsible for the inferior performances when using the statistical strategy. The procedure of model selection is confined to comparisons between models with a nested relationship, thereby impairing the performance of the statistical strategy.

B. FUTURE WORKS

In this study we have focused on a general regression model paradigm that provides a natural solution from two part form of MDL to the GCA. As a general principle for statistical model selection, the MDL principle develops many forms of description length in terms of coding schemes. The MDL forms can be viewed as imposing an adaptive penalty on model size. Although all forms achieve pointwise and minimax lower bounds on redundancy [72], further investigation is required to determine the optimal coding scheme suitable for a given neuroimaging modality or noise level. Future animal experiments that provide intracranial recordings of and fMRI measurements of neural network synchronously on the same animal will help to direct and test the development of these forms. Moreover, the GCA schemes were also conveyed in different function spaces [49], [89]. The MDL will provide potential approaches to the Granger-causality representations in other function spaces principle because has rich connections with Bayesian statistics [72]. More importantly, the MDL principle allows the comparison between any two model classes in terms of code length, regardless of their forms [72]. This robust feature has potential to accommodate the representations of Granger-causality generated from different model classes, while the causal correlation can be also investigated between disparate function spaces.

REFERENCES

[1] K. Amunts, C. Ebell, J. Müller, M. Telefont, A. Knoll, and T. Lippert, "The human brain project: Creating a European research infrastructure to decode the human brain," *Neuron*, vol. 92, no. 3, pp. 574–581, Nov. 2016.

[2] C. L. Martin and M. Chun, "The BRAIN initiative: Building, strengthening, and sustaining," *Neuron*, vol. 92, no. 3, pp. 570–573, Nov. 2016.

[3] M.-M. Poo, J.-L. Du, N. Y. Ip, Z.-Q. Xiong, B. Xu, and T. Tan, "China brain project: Basic neuroscience, brain diseases, and brain-inspired computing," *Neuron*, vol. 92, no. 3, pp. 591–596, Nov. 2016.

[4] A. B. A. S. Committee, "Australian brain alliance," *Neuron*, vol. 92, no. 3, pp. 597–600, Nov. 2016.

[5] H. Okano, E. Sasaki, T. Yamamori, A. Iriki, T. Shimogori, Y. Yamaguchi, K. Kasai, and A. Miyawaki, "Brain/MINDS: A japanese national brain project for marmoset neuroscience," *Neuron*, vol. 92, no. 3, pp. 582–590, Nov. 2016.

[6] S.-J. Jeong, H. Lee, E.-M. Hur, Y. Choe, J. W. Koo, J.-C. Rah, K. J. Lee, H.-H. Lim, W. Sun, C. Moon, and K. Kim, "Korea brain initiative: Integration and control of brain functions," *Neuron*, vol. 92, no. 3, pp. 607–611, Nov. 2016.

[7] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, Mar. 2001.

[8] O. Sporns, D. Chialvo, M. Kaiser, and C. Hilgetag, "Organization, development and function of complex brain networks," *Trends Cognit. Sci.*, vol. 8, no. 9, pp. 418–425, Sep. 2004.

[9] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nature Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, Mar. 2009.

[10] S. L. Bressler and V. Menon, "Large-scale brain networks in cognition: Emerging methods and principles," *Trends Cognit. Sci.*, vol. 14, no. 6, pp. 277–290, Jun. 2010.

[11] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, Sep. 2010.

[12] O. Sporns, "The human connectome: A complex network," *Ann. New York Acad. Sci.*, vol. 1224, no. 1, pp. 109–125, Apr. 2011.

[13] M. Siegel, T. H. Donner, and A. K. Engel, "Spectral fingerprints of large-scale neuronal interactions," *Nature Rev. Neurosci.*, vol. 13, no. 2, pp. 121–134, Feb. 2012.

[14] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. D. Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo, D. A. Handwerker, S. Keilholz, V. Kiviniemi, D. A. Leopold, F. de Pasquale, O. Sporns, M. Walter, and C. Chang, "Dynamic functional connectivity: Promise, issues, and interpretations," *NeuroImage*, vol. 80, pp. 360–378, Oct. 2013.

[15] S. L. Valk, B. C. Bernhardt, A. Böckler, F.-M. Trautwein, P. Kanske, and T. Singer, "Socio-cognitive phenotypes differentially modulate large-scale structural covariance networks," *Cerebral Cortex*, vol. 27, Jan. 2016, Art. no. bhv319.

[16] K. E. Stephan and A. Roebroeck, "A short history of causal modeling of fMRI data," *NeuroImage*, vol. 62, no. 2, pp. 856–863, Aug. 2012.

[17] J. Tsunada, A. S. K. Liu, J. I. Gold, and Y. E. Cohen, "Causal contribution of primate auditory cortex to auditory perceptual decision-making," *Nature Neurosci.*, vol. 19, no. 1, pp. 135–142, Jan. 2016.

[18] O. David, I. Guillemain, S. Saillet, S. Reyt, C. Deransart, C. Segebarth, and A. Depaulis, "Identifying neural drivers with functional MRI: An electrophysiological validation," *PLoS Biol.*, vol. 6, no. 12, p. e315, 2008.

[19] T. Van Kerkoerle, M. W. Self, B. Dagnino, M.-A. Gariel-Mathis, J. Poort, C. Van Der Togt, and P. R. Roelfsema, "Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 40, pp. 14332–14341, 2014.

[20] A. M. Bastos, J. Vezoli, C. A. Bosman, J.-M. Schoffelen, R. Oostenveld, J. R. Dowdall, P. De Weerd, H. Kennedy, and P. Fries, "Visual areas exert feedforward and feedback influences through distinct frequency channels," *Neuron*, vol. 85, no. 2, pp. 390–401, Jan. 2015.

[21] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, "Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 26, pp. 9849–9854, Jun. 2004.

[22] X. Wang, Y. Chen, and M. Ding, "Estimating granger causality after stimulus onset: A cautionary note," *NeuroImage*, vol. 41, no. 3, pp. 767–776, Jul. 2008.

[23] D. W. Gow, J. A. Segawa, S. P. Ahlfors, and F.-H. Lin, "Lexical influences on speech perception: A granger causality analysis of MEG and EEG source estimates," *NeuroImage*, vol. 43, no. 3, pp. 614–623, Nov. 2008.

[24] M. Ploner, J.-M. Schoffelen, A. Schnitzler, and J. Gross, "Functional integration within the human pain system as revealed by granger causality," *Hum. Brain Mapping*, vol. 30, no. 12, pp. 4025–4032, Dec. 2009.

[25] E. Florin, J. Gross, J. Pfeifer, G. R. Fink, and L. Timmermann, "The effect of filtering on granger causality based multivariate causality measures," *NeuroImage*, vol. 50, no. 2, pp. 577–588, Apr. 2010.

[26] A. Roebroeck, E. Formisano, and R. Goebel, "Mapping directed influence over the brain using granger causality and fMRI," *NeuroImage*, vol. 25, no. 1, pp. 230–242, Mar. 2005.

[27] D. Sridharan, D. J. Levitin, and V. Menon, "A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 34, pp. 12569–12574, Aug. 2008.

[28] V. Menon and L. Q. Uddin, "Saliency, switching, attention and control: A network model of insula function," *Brain Struct. Function*, vol. 214, nos. 5–6, pp. 655–667, Jun. 2010.

[29] S. Ryali, K. Supekar, T. Chen, and V. Menon, "Multivariate dynamical systems models for estimating causal interactions in fMRI," *NeuroImage*, vol. 54, no. 2, pp. 807–823, Jan. 2011.

[30] N. Wiener, "The theory of prediction," in *Modern Mathematics for Engineers*. New York, NY, USA: McGraw-Hill, 1956.

[31] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, p. 424, Aug. 1969.

[32] C. Granger and P. Newbold, "Spurious regressions in econometrics.," *J. Econometrics*, vol. 2, no. 2, pp. 111–120, 1974.

[33] M. Chávez, J. Martinerie, and M. Le Van Quyen, "Statistical assessment of nonlinear causality: Application to epileptic EEG signals," *J. Neurosci. Methods*, vol. 124, no. 2, pp. 113–128, Apr. 2003.

[34] B. Gourévitch, R. L. Bouquin-Jeannès, and G. Faucon, "Linear and nonlinear causality between signals: Methods, examples and neurophysiological applications," *Biol. Cybern.*, vol. 95, no. 4, pp. 349–369, Sep. 2006.

[35] R. Goebel, A. Roebroeck, D.-S. Kim, and E. Formisano, "Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and granger causality mapping," *Magn. Reson. Imag.*, vol. 21, no. 10, pp. 1251–1261, Dec. 2003.

[36] R. S. Hacker and A. Hatemi-J, "Tests for causality between integrated variables using asymptotic and bootstrap distributions: Theory and application," *Appl. Econ.*, vol. 38, no. 13, pp. 1489–1500, Jul. 2006.

[37] M. Kaminski, *Multichannel Data Analysis in Biomedical Research*, 2nd ed. Pacific Grove, CA, USA: Duxbury Resource Center, 2007.

[38] A. Hatemi-J, "Asymmetric causality tests with an application," *Empirical Econ.*, vol. 43, no. 1, pp. 447–456, Aug. 2012.

[39] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *J. Amer. Stat. Assoc.*, vol. 77, no. 378, pp. 304–313, Jun. 1982.

[40] M. Ding, Y. Chen, and S. L. Bressler, "Granger causality: Basic theory and application to neuroscience," *Quant. Biol.*, vol. 437, pp. 826–831, Dec. 2006.

[41] S. Guo, J. Wu, M. Ding, and J. Feng, "Uncovering interactions in the frequency domain," *PLoS Comput. Biol.*, vol. 4, no. 5, 2008, Art. no. e1000087.

[42] A. K. Seth, "A MATLAB toolbox for granger causal connectivity analysis," *J. Neurosci. Methods*, vol. 186, no. 2, pp. 262–273, Feb. 2010.

[43] L. Barnett and A. K. Seth, "The MVGC multivariate granger causality toolbox: A new approach to Granger-causal inference," *J. Neurosci. Methods*, vol. 223, pp. 50–68, Feb. 2014.

[44] P. A. Stokes and P. L. Purdon, "A study of problems encountered in granger causality analysis from a neuroscience perspective," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 34, pp. E7063–E7072, Aug. 2017.

[45] L. Ning and Y. Rathi, "A dynamic regression approach for frequency-domain partial coherence and causality analysis of functional brain networks," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 1957–1969, Sep. 2018.

[46] W. Liao, D. Marinazzo, Z. Pan, Q. Gong, and H. Chen, "Kernel Granger causality mapping effective connectivity on fMRI data," *IEEE Trans. Med. Imag.*, vol. 28, no. 11, pp. 1825–1835, Nov. 2009.

[47] Z. Chen, K. Zhang, L. Chan, and B. Schölkopf, "Causal discovery via reproducing kernel Hilbert space embeddings," *Neural Comput.*, vol. 26, no. 7, pp. 1484–1517, Jul. 2014.

[48] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Feb. 2004, Art. no. 026113.

[49] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear granger causality," *Phys. Rev. Lett.*, vol. 100, no. 14, Apr. 2008, Art. no. 144103.

[50] C. Zou, C. Ladroue, S. Guo, and J. Feng, "Identifying interactions in the time and frequency domains in local and global networks–A granger causality approach," *BMC Bioinf.*, vol. 11, no. 1, p. 337, 2010.

[51] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, 1974.

[52] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[53] G. Deshpande, S. LaConte, G. A. James, S. Peltier, and X. Hu, "Multivariate granger causality analysis of fMRI data," *Human Brain Mapping*, vol. 30, no. 4, pp. 1361–1373, Apr. 2009.

[54] S. Guo, C. Ladroue, and J. Feng, "Granger causality: Theory and applications," in *Frontiers in Computational and Systems Biology*. London, U.K.: Springer, 2010, pp. 83–111.

[55] S. L. Bressler and A. K. Seth, "Wiener–Granger causality: A well established methodology," *NeuroImage*, vol. 58, no. 2, pp. 323–329, Sep. 2011.

[56] X. Li, G. Marrelec, R. F. Hess, and H. Benali, "A nonlinear identification method to study effective connectivity in functional MRI," *Med. Image Anal.*, vol. 14, no. 1, pp. 30–38, Feb. 2010.

[57] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA, USA: Duxbury Resource Center, 2001.

[58] R. L. Wasserstein and N. A. Lazar, "The ASA statement on *p*-values: Context, process, and purpose," *Amer. Statistician*, vol. 70, no. 2, pp. 129–133, 2016.

[59] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar, "Moving to a world beyond '*p*< 0.05,'" *Amer. Statistician*, vol. 73, pp. 1–19, Mar. 2019.

[60] V. Amrhein, S. Greenland, and B. McShane, "Scientists rise up against statistical significance," *Nature*, vol. 567, no. 7748, pp. 305–307, Mar. 2019.

[61] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett, "Abandon statistical significance," *Amer. Statistician*, vol. 73, pp. 235–245, Mar. 2019.

[62] V. Amrhein and S. Greenland, "Remove, rather than redefine, statistical significance," *Nature Hum. Behav.*, vol. 2, no. 1, p. 4, Jan. 2018.

[63] V. Amrhein, D. Trafimow, and S. Greenland, "Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication," *Amer. Statistician*, vol. 73, pp. 262–270, Mar. 2019.

[64] S. H. Hurlbert, R. A. Levine, and J. Utts, "Coup de grâce for a tough old bull: 'Statistically Significant' expires," *Amer. Statistician*, vol. 73, pp. 352–357, Mar. 2019.

[65] X. Miao, X. Wu, R. Li, K. Chen, and L. Yao, "Altered connectivity pattern of hubs in default-mode network with Alzheimer's disease: An granger causality modeling approach," *PLoS ONE*, vol. 6, no. 10, 2011, Art. no. e25546.

[66] G. Deshpande, P. Santhanam, and X. Hu, "Instantaneous and causal connectivity in resting state brain networks derived from functional MRI data," *NeuroImage*, vol. 54, no. 2, pp. 1043–1052, Jan. 2011.

[67] X. Li, D. Coyle, L. Maguire, T. M. McGinnity, and H. Benali, "A model selection method for nonlinear system identification based fMRI effective connectivity analysis," *IEEE Trans. Med. Imag.*, vol. 30, no. 7, pp. 1365–1380, Jul. 2011.

[68] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.

[69] P. G. Bryant and O. I. Cordero-Brana, "Model selection using the minimum description length principle," *Amer. Statistician*, vol. 54, no. 4, p. 257, Nov. 2000.

[70] P. D. Grünwald, I. J. Myung, and M. A. Pitt, "Advances in Minimum Description Length: Theory and Applications," in *Neural Information Processing Series*. Cambridge, MA, USA: MIT Press, 2005.

[71] M. Hansen and B. Yu, "Bridging aic and bic: An MDL model selection criterion," in *Proc. Inf. Theory Workshop Detection, Estimation, Classification Imag.*, Santa Fe, NM, USA, Feb. 1999, pp. 24–26.

[72] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Publications Amer. Stat. Assoc.*, vol. 96, no. 454, pp. 746–774, 2001.

[73] M. Li *et al.*, *An Introduction to Kolmogorov Complexity and Its Applications*, vol. 3. New York, NY, USA: Springer, 2008.

[74] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.

[75] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, no. 2, pp. 416–431, Jun. 1983.

[76] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 4, pp. 629–636, Jul. 1984.

[77] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, Sep. 1986.

[78] J. Rissanen, "Stochastic complexity," *J. Roy. Stat. Soc., Ser. B, (Methodol.)*, vol. 49, no. 3, pp. 223–239, 1987.

[79] J. J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 40–47, Jan. 1996.

[80] A. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems Inf. Transmiss.*, vol. 1, no. 1, pp. 1–7, 1965.

[81] A. Kolmogorov, "Logical basis for information theory and probability theory," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 5, pp. 662–664, Sep. 1968.

[82] C. S. Wallace and D. M. Boulton, "An information measure for classification," *Comput. J.*, vol. 11, no. 2, pp. 185–194, Aug. 1968.

[83] J. Rissanen, "Information and complexity in statistical modeling," in *Information Science and Statistics*. New York, NY, USA: Springer, 2005.

[84] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[85] P. D. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, USA: MIT Press, 2007.

[86] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.

[87] F. Li, X. Wang, P. Shi, Q. Lin, and Z. Hu, "Neural network can be revealed with functional MRI: Evidence from self-consistent experiment," *Nature Commun.*, to be published.

[88] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston, ''Effective connectivity: Influence, causality and biophysical modeling,'' *NeuroImage*, vol. 58, no. 2, pp. 339–361, Sep. 2011.

[89] L. Angelini, M. Pellicoro, and S. Stramaglia, ''Granger causality for circular variables,'' *Phys. Lett. A*, vol. 373, no. 29, pp. 2467–2470, Jun. 2009.

FEI LI was born in Hubei, China, in 1995. He received the M.S. degree from the College of Science, Zhejiang University of Technology, Hangzhou, Zhejiang, China, in 2017, where he is currently pursuing the Ph.D. degree.

XUEWEI WANG was born in Shanxi, China, in 1997. She received the M.S. degree from the College of Science, Zhejiang University of Technology, Hangzhou, Zhejiang, China, in 2018.

QIANG LIN was born in Zhejiang, China, in July 1964. He received the M.S. degree in optics from Hangzhou University, in 1988, and the Ph.D. degree from Zhejiang University, in 2003. He was a Professor and the Head of the Optics Division with the Department of Physics, Hangzhou University, in 1994. From 1996 to 1997, he was with the Technical University of Vienna, Austria. In 1998, he joined the Department of Physics, Zhejiang University, and became the Director of the Optics Institute. From 2005 to 2006, he was an Alexander von Humboldt Fellow with the Max-Born-Institute, Berlin, Germany. Since 2014, he was the Dean of the College of Sciences, Zhejiang University of Technology. He is the author or coauthor of more than 200 scientific articles and three books.

ZHENGHUI HU received the Ph.D. degree in physics from Zhejiang University, Hangzhou, China, in 2005. From 2005 to 2007, he was a Postdoctoral Fellow in electrical and computer engineering with The Hong Kong University of Science and Technology. From 2008 to 2010, he was a Postdoctoral Fellow in computing and information science with the Rochester Institute of Technology, Rochester, NY, USA. Since 2015, he has been a Professor with the College of Science, Zhejiang University of Technology, Hangzhou. His research interests include functional brain imaging, magnetocardiography, and signal reconstruction.

● ● ●