Jacobian Regularizer-based Neural Granger Causality

Wanqi Zhou [1 2] **Shuanghao Bai** [1] **Shujian Yu** [3] **Qibin Zhao** [2] **Badong Chen** [1]

Abstract

With the advancement of neural networks, diverse methods for neural Granger causality have emerged, which demonstrate proficiency in handling complex data, and nonlinear relationships. However, the existing framework of neural Granger causality has several limitations. It requires the construction of separate predictive models for each target variable, and the relationship depends on the sparsity on the weights of the first layer, resulting in challenges in effectively modeling complex relationships between variables as well as unsatisfied estimation accuracy of Granger causality. Moreover, most of them cannot grasp full-time Granger causality. To address these drawbacks, we propose a **J**acobian **R**egularizer-based **N**eural **G**ranger **C**ausality (**JRNGC**) approach, a straightforward yet highly effective method for learning multivariate summary Granger causality and full-time Granger causality by constructing a single model for all target variables. Specifically, our method eliminates the sparsity constraints of weights by leveraging an input-output Jacobian matrix regularizer, which can be subsequently represented as the weighted causal matrix in the post-hoc analysis. Extensive experiments show that our proposed approach achieves competitive performance with the state-of-the-art methods for learning summary Granger causality and full-time Granger causality while maintaining lower model complexity and high scalability.

1. Introduction

On time-series data, causal relations can be learned by various methodologies, such as Granger causality (Granger,

1969), transfer entropy (TE) (Schreiber, 2000), and other approaches like constraint-based methods (e.g., PCMCI), noise-based methods (e.g., VarLiNGAM), and more, as discussed in the literature (Assaad et al., 2022). Traditionally, methodologies like conditional TE (Sun et al., 2015; Zhou et al., 2022) and constraint-based algorithms such as PCMCI, tsFCI (Runge et al., 2019; Entner & Hoyer, 2010) have been employed for uncovering nonlinear causal relationships. These methods have become popular due to their capacity to sidestep the necessity for parametric modeling. However, they grapple with a significant challenge when it comes to scaling up to large causal graphs because of their iterative nature and dependence on the estimation of probability densities, which motivates the need for more efficient and effective techniques in causal relationship identification within time series datasets.

Recent advancements in deep learning have shown promising progress in learning Granger causality. Several studies have explored the application of neural networks, such as multi-layer perceptron (MLP), recurrent neural networks (RNNs), and dynamic variational autoencoders, for learning Granger causality from time series data (Tank et al., 2021; Suryadi et al., 2023; Bussmann et al., 2021; Li et al., 2023). These methods integrate Granger causality as a null hypothesis into time series forecasting or generative models, allowing causal relationships between variables to be more accurately captured through causal constraint modeling. Despite these advancements, current neural Granger causality methods have two major drawbacks. Firstly, as illustrated in the left panel of Figure 1, they require building separate neural networks for each target variable during the training phase to learn Granger causality. This can result in computational inefficiencies and scalability challenges when dealing with larger and more complex datasets, especially when we need to grasp long full-time Granger causality. Secondly, they rely on sparse constraints on the first layer weights, which leads to the necessity of separate models for target variables, and also limits the ability to capture Granger causal relationships for complex data, because the dependencies between variables over time will be overlooked. Moreover, in cases where the number of variables is high, the search space for causal variance becomes broad. Thus solely constraining the first layer is insufficient and may result in local optima. In other words, the relationship between
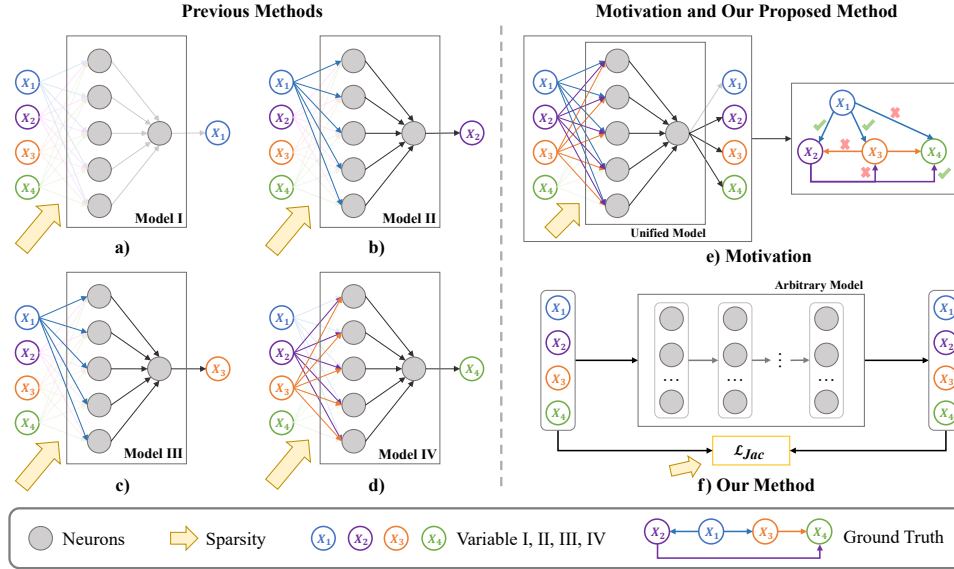
---

[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China. [2]RIKEN AIP, Japan. [3]Vrije Universiteit Amsterdam, Netherlands. Correspondence to: Badong Chen <chenbd@mail.xjtu.edu.cn>.

1

Figure 1. Motivation and our proposed neural Granger causality method. To learn the true Granger causality, we need to estimate the importance of a variable in helping forecast another variable. For example, let's examine a simple scenario involving the summary causal relationship: $X_4 \leftarrow X_2 \leftarrow X_1 \rightarrow X_3 \rightarrow X_4$. To comprehend this relationship, current neural Granger causality methods need to construct and train the same number of models as the dimensions of the variables to disentangle the importance of each variable and obtain Granger causality by incorporating sparsity penalties on the first layer of each model, as illustrated in figures a)-d). However, sparse first-layer network parameters will result in challenges in effectively modeling complex relationships between variables as well as unsatisfied estimation accuracy of Granger causality. In addition, it will get wrong Granger causality if we use a multivariate time series forecasting model because of the existence of a shared hidden layer, as exemplified in figure e). Instead, our method only needs to build and train a single multivariate time series forecasting model by introducing input-output Jacobian regularizer $\mathcal{L}_{Jac}$. Note that the numbers I, II, III, and IV mean the same model architecture with independent training for different variables and we used two-layer Perceptron for the convenience of illustration.

predicting variables and learning causal connections is mutually influential; accurate predictions facilitate genuine causal learning, and, conversely, a proficient understanding of causal relationships enhances predictive accuracy. Furthermore, these methods render RNNs (e.g., RNN, LSTM) incapable of learning full-time Granger causality, because in a standard RNN, different time steps of the same time series share the same weight parameters, thus preventing the model from capturing different Granger causality across time steps. This limitation compromises the full realization of neural networks' potential in certain applications.

For the first issue, to develop a unified neural Granger causality framework with a multivariate time series forecasting model, it is essential to mitigate the impact of shared hidden layers. In the context of a multivariate time series forecasting model, shared hidden layers would obtain the information from all variables and send it to each target variable, leading to the difficulty of disentangling the importance of each variable. Thus we must identify a suitable Granger causality estimation method capable of expressing the importance of one variable to another. For the second issue, we need to find a global Granger causality constraint method. To solve these issues, we propose a novel approach

termed the **J**acobian **R**egularization-based **N**eural **G**ranger **C**ausality (JRNGC). For a detailed view, see the right panel in Figure 1. Firstly, the input-output Jacobian matrix can represent the relations between variables over time and does not purely rely on the first layer of the neural network. By imposing a sparse constraint on it, we can explicitly set the null hypothesis that there is no Granger causality between certain variables. Therefore, the input-output Jacobian matrix can handle the above-mentioned two major drawbacks of NGC. Furthermore, we utilize the residual MLP neural network, proven more capable of predicting time series (Das et al., 2023; Zeng et al., 2023), although our framework is adaptable to any model.

Our main contributions to this work are as follows:

- To our best knowledge, this is the first work to harness a single NN model with shared hidden layers for multivariate Granger causality analysis.

- We propose a novel neural network framework to learn Granger causality by incorporating an input-output Jacobian regularizer in the training objective.

- Our method can not only obtain the summary Granger

causality but also the full-time Granger causality.

- We evaluate our method on commonly used benchmark datasets with extensive experiments. Our method can outperform state-of-the-art baselines and show an excellent ability to discover Granger causality, especially for sparse causality.

2. Background and Related Work

2.1. Background: Granger causality

Granger causality (Granger, 1969) is widely used for learning causality from time-series data. It is a statistical concept of causality that is based on prediction. It follows Wiener's theory (Wiener, 1956): if the prediction of variable B can be improved by incorporating the past information of variable A, along with its own past information, then A Granger causes B. Given a multivariate time series $\mathbf{x} = \{x_1, x_2, \cdots, x_D\}$ with D dimensions, we model the time series as:

$$x_j(t) = f_j\left(x_1(< t), \ldots, x_D(< t)\right) + \varepsilon_j, \qquad (1)$$

where the ε_j is an independent noise item, and $x_i(< t)$ denotes the past information of time series x_i. The prediction of the value of x_j at time t depends on the past information of other time series, which are the potential causes or "parent" of x_j. If a time series x_i (at times before t, i.e., $x_i(< t)$) significantly improves the accuracy of predicting the future values of x_j when added to the set of "parent" time series, then x_i is said to have Granger causality on x_j.

The definition of the summary causal graph and full-time causal graph can be seen in the appendix.

2.2. Related Work

Traditionally, the Granger causality test involved comparing prediction errors from linear models with and without past information for two variables to assess their causal relationship through hypothesis testing. However, this approach, based on linear Vector Autoregressive (VAR) had limitations in capturing all types of interactions. While some methods attempted to use kernel techniques to handle nonlinear systems (Marinazzo et al., 2008; Amblard et al., 2012), they still lacked the flexibility of non-parametric models like transfer entropy and its invariants (Zhou et al., 2022). Nonetheless, these non-parametric methods often require large amounts of data and suffer from slow computation, among other issues. In recent years, machine learning, especially deep neural networks has been leveraged to enhance the capture of Granger causality, especially in dealing with nonlinear relationships and complex data. As highlighted in (Tank et al., 2021), integrating neural networks into the framework of Granger causality faces two key challenges: the lack of

interpretability and the difficulty in disentangling the importance of variables in the presence of shared hidden layers for a joint network. This complicates the specification of selective Granger causality conditions. To tackle these challenges, (Tank et al., 2021) chose to individually model each output component, apply a group sparse group lasso penalty on the weights of the first layer, and optimize by proximal gradient descent.

Most neural Granger Causality approaches follow the idea in (Tank et al., 2021), utilizing various time series forecasting models to improve the interpretability of causal relations learned by black-box methods. For instance, (Khanna & Tan, 2020) employed an economy statistical recurrent units (eSRU) model, while (Bussmann et al., 2021) proposed an additive neural Vector Autoregressive (VAR) model. Additionally, (Suryadi et al., 2023) modified the architecture by transforming fully connected hidden layers from the input layer into a one-to-one configuration. Alternatively, (Li et al., 2023) utilizes a dynamic variational autoencoder model, which learns causality and enhances generative performance. Concerning the causal sparsity constraint, all mentioned methods enforce sparsity either on the weights of the first layer or the weights of the final layer.

Based on our current understanding, we observe that existing methods predominantly rely on single-variable time prediction models. This raises the following considerations:

Q1: *Why do these methods primarily rely on single-variable prediction models? Why not utilize multivariate prediction models?*

A1: These methods depend on neural network weights, either in the first layer or the last layer, to express and constrain Granger causality. Utilizing single-variable neural network prediction models aids in decoupling the relationships between variables, whereas in multivariate prediction models, the shared hidden layer prevents the isolation of variable importance, as exemplified in Figure 1.

Q2: *What are the existing issues with current methods?*

A2: These methods restrict the neural network's ability to learn complex data relationships and additional computation. In particular, RNNs, which are suitable for non-stationary time series, fail to capture full-time Granger causality in current framework.

Q3: *How can we overcome the aforementioned challenges?*

A3: We propose constraining the Jacobian matrix between inputs and outputs. This approach eliminates the need to express Granger causality through network weights, enabling the learning of summary and full-time Granger causality on any multivariate time series prediction network that is suitable for the data characteristics.

3. Method

Given a D-dimensional multivariate time series $\mathbf{x} = \{x_1, x_2, \dots, x_D\}$, as shown in Figure 1, our model can be summarized as follows:

1. With the past information of each time series $\mathbf{x}_{1:D}^{t-\tau:t-1}$ as input, we forecast their future $\mathbf{x}_{1:D}^{t}$ by a multivariate time series forecasting model. In this work, we leverage a residual MLP network.

2. In the training phase, we train the model with mean squared error loss along with the input-output Jacobian matrix regularizer, which facilitates the model's learning of both Granger causality and forecasting.

3. In the post-hoc analysis phase, we use the learned input-output Jacobian matrix to analyze the summary and full-time Granger causality between time series.

In the following part, we will provide detailed descriptions of the proposed model, regularizer term, and post-hoc analysis method. Before that, it is crucial to emphasize that we leverage the residual MLP network for its advantages in forecasting time series (Das et al., 2023; Zeng et al., 2023); however, our proposed framework is versatile and can be applied to alternative models, such as LSTM (see the Experiments section). With this approach, we can also achieve full-time Granger causality.

3.1. Residual MLP-based time series forecasting model

Granger causality relies on prediction as a fundamental component. It always uses the time series forecasting model to discover the Granger causality. A good predictive model with high performance, free from overfitting, and resilient to noise, can facilitate Granger causality learning. Due to the success of residual MLP-based models in time series analysis (Das et al., 2023; Zeng et al., 2023), which are employed in state-of-the-art (SOTA) methods for predicting time series, we utilize a simple MLP-based model to learn Granger causality. Our model includes an input layer and an individual output layer, along with a varying number of robust residual MLP layers that seamlessly join together to create a unified framework for analyzing multivariate time series data, which can be formulated as follows:

$$\hat{\mathbf{x}}_{1:D}^{t} = \text{FC}_2(\text{ResidualBlock}(\text{FC}_1(\mathbf{x}_{1:D}^{t-\tau:t-1}))), \quad (2)$$

where FC_1 and FC_2 represent fully-connected layers. ResidualBlock is composed of n residual MLP layers, as shown in the Figure 2. Since this architecture does not have any self-attention, recurrent, or convolutional mechanisms, it allows our model to benefit from the simplicity and speed of linear models while still being able to handle non-linear dependencies.
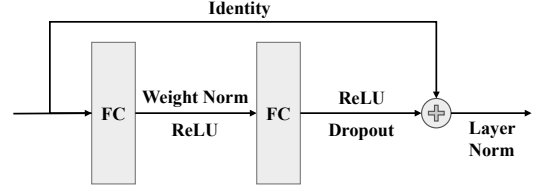

Figure 2. The framework of the Residual MLP layer in this work. FC represents the fully connected layer.

3.2. Input-output Jacobian regularizer as a Granger causality constraint

It is necessary to incorporate the null hypothesis that the variable x_i is not the Granger causality of x_j as a Granger causality constraint within the model. Conventionally, this is accomplished by applying L_1 or L_2 norm regularizer to weights of the first layer (Tank et al., 2021; Suryadi et al., 2023). This regularizer approach contributes to learning the Granger causal connections between variables in the case of univariate time series prediction, where the weight can be expressed as a direct relationship from one variable to another. However, this principle encounters three challenges. Firstly, imposing a penalty on the first layer may hinder the neural network from adequately learning the time series. Secondly, the weights of the first layer prevent RNNs, LSTMs, and other recurrent networks from learning full-time Granger causality. Thirdly, if one wishes to alter the current framework for learning Granger causality in multivariate multi-model settings, it becomes necessary to abandon this approach. In contrast, the input-output Jacobian matrix allows us to find the relationships between variables, including their interactions, self-importance, and dependencies over time. The definition of the Jacobian matrix is as follows:

$$\mathbf{J} = \left[\begin{array}{ccc} \frac{\partial \mathbf{f}}{\partial x_1^{t-\tau}} & \cdots & \frac{\partial \mathbf{f}}{\partial x_D^{t-1}} \end{array} \right]$$
$$= \left[\begin{array}{ccccccc} \frac{\partial f_1}{\partial x_1^{t-\tau}} & \cdots & \frac{\partial f_1}{\partial x_1^{t-1}} & \cdots & \frac{\partial f_1}{\partial x_D^{t-\tau}} & \cdots & \frac{\partial f_1}{\partial x_D^{t-1}} \\ & & & \vdots & & & \\ \frac{\partial f_D}{\partial x_1^{t-\tau}} & \cdots & \frac{\partial f_D}{\partial x_1^{t-1}} & \cdots & \frac{\partial f_D}{\partial x_D^{t-\tau}} & \cdots & \frac{\partial f_D}{\partial x_D^{t-1}} \end{array} \right], \quad (3)$$

where the f_i represents the i-th predictive function for $i-$th variable.

To detect the lags where Granger causal effects exist, we enforce the sparsity of the input-output Jacobian matrix. Nevertheless, the complexity of computing the L_1 norm of a Jacobian matrix of size $(D, D\tau)$ is $O(g(D, \tau))$, and $g(D, \tau)$ represents the specific time and resource requirements of the algorithm. It will cost lots of computing time when the D is high, which can be seen in Table 12.

Instead, we regularize the squared Frobienus norm of the input-output Jacobian matrix as (Hoffman et al., 2019),

which can be efficiently computed:

$$\|J(\boldsymbol{x})\|_{\mathrm{F}}^2 = \mathrm{Tr}(JJ^{\mathrm{T}}) = \sum_{\{e\}} eJJ^{\mathrm{T}}e^{\mathrm{T}}$$
$$= \sum_{\{e\}} \left[\frac{\partial(\boldsymbol{e} \cdot \boldsymbol{z})}{\partial \boldsymbol{x}} \right]^2, \quad (4)$$

where $\{e\}$ denotes a constant orthonormal basis of the D-dimensional output space. $\boldsymbol{z}$ is the output with respect to input variables $\boldsymbol{x}$. $\mathrm{Tr}(\cdot)$ represents the trace function. Ultimately, this leads to computational overhead that increases linearly with the output dimension D.

As illustrated in (Hoffman et al., 2019), we can rewrite the Eq. 4 and use random projection to compute the squared Frobienus norm of the input-output Jacobian matrix efficiently, which projects the high-dimensional data onto a lower-dimensional space, thereby reducing computational and storage costs:

$$\|J(\boldsymbol{x})\|_{\mathrm{F}}^2 \approx \frac{1}{n_{\mathrm{proj}}} \sum_{\mu=1}^{n_{\mathrm{proj}}} \left[\frac{\partial(\hat{\boldsymbol{v}}^{\mu} \cdot \boldsymbol{z})}{\partial \boldsymbol{x}} \right]^2, \quad (5)$$

where the random vector $\hat{\boldsymbol{v}}^{\mu}$ is drawn from the $(D-1)$-dimensional unit sphere S^{D-1}, n_{proj} is the number of random projection. Utilizing a mini-batch size of $|B| = 100$, a singular projection results in model performance nearly indistinguishable from the exact method, while significantly reducing computational costs by orders of magnitude.

Above all, our method can be named JRNGC-L1, JRNGC-F according to different norm regularizers. The penalized loss function of JRNGC-L1 can be formulated as follows:

$$\frac{1}{D} \sum_{i=1}^{D} \frac{1}{N-\tau} \sum_{t=\tau+1}^{N} (\hat{x}_{i,t} - x_{i,t})^2 + \lambda \|J(\boldsymbol{x})\|_1. \quad (6)$$

The penalized loss function of JRNGC-F can be formulated as follows:

$$\frac{1}{D} \sum_{i=1}^{D} \frac{1}{N-\tau} \sum_{t=\tau+1}^{N} (\hat{x}_{i,t} - x_{i,t})^2 + \lambda \|J(\boldsymbol{x})\|_{\mathrm{F}}^2, \quad (7)$$

where, in both Eq. 6 and Eq. 7, the λ mean the regularizer coefficient and D is the dimension of time series and N is the length of time. Note that, although the Jacobian regularizer is popular in computer vision (Jakubovitz & Giryes, 2018; Rhodes & Lee, 2021), to our best knowledge, we are the first to use it in Granger causality learning.

3.3. Input-output Jacobian matrix as the variable's causal importance

In the post-hoc analysis, we employ the learned input-output Jacobian matrix to analyze the causal relationships between variables. The input-output Jacobian matrix can obtain the variable's causal importance through $J_{i,j,\alpha} = \frac{\partial x_j}{\partial x_i^{(t-\alpha)}}$. It allows us to obtain the variable's lag importance, i.e., the full-time Granger causal graph.

4. Experiments

In this section, we demonstrate the performance of the proposed methods, i.e., JRNGC-L1, JRNGC-F on five widely used benchmarks: the VAR model, the Lorenz-96 model, fMRI data, the DREAM-3 dataset and CausalTime (Cheng et al., 2024b). We perform comparative experiments with competitive methods, including GC (Granger, 1969), PCMCI (Runge et al., 2019), cMLP (Tank et al., 2021), cLSTM (Tank et al., 2021), NAVAR (MLP) (Bussmann et al., 2021), NAVAR (LSTM) (Bussmann et al., 2021), SRU, eSRU(Khanna & Tan, 2020), TCDF (Nauta et al., 2019), JGC (Suryadi et al., 2023) and CR-VAE (Li et al., 2023), Scalable Causal Graph Learning (SCGL, (Xu et al., 2019)), CUTS (Cheng et al., 2023), CUTS+ (Cheng et al., 2024a), NTS-NOTEARS (abbreviated as N.NTS, (Sun et al., 2023)), Rhino (Gong et al., 2023), and Latent Convergent Cross Mapping (LCCM, (Brouwer et al., 2021)), as well as Neural Graphical Model (NGM, (Bellot et al., 2022)). Additionally, our summary hyperparameters of experiments for all models and other additional experiments are detailed in the appendix.

4.1. Metrics

We employ two standard metrics: the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). A value of 0.5 or lower in AUROC indicates poor performance. In scenarios with sparse causal relationships, AUPRC becomes a more reliable indicator of the model's ability to detect causal edges. This is due to its emphasis on correctly identifying positive instances, a critical aspect when the number of actual causal relationships is limited.

4.2. Experiment results and analysis

In this section, we demonstrate the effectiveness of our proposed method in inferring Granger causality between variables, i.e., the summary Granger causality, and the full-time Granger causality. Specifically, we use the term "with no lag" to express Granger causality between variables, and "with lag" to express full-time Granger causality. Our results show that our proposed method is capable of learning summary Granger causality and full-time Granger causality with high AUROC and AUPRC scores on both synthetic and open-real benchmark datasets.

VAR model. For this model, we simulated up to $T = 600$ observations with the maximum true time lag $\tau \in \{3, 5\}$,

Table 1. Comparisons of AUROC and AUPRC for Granger causality among different approaches on VAR dataset. The mean results with standard error are reported by averaging over 5 runs. Note that "with lag" refers to the full-time causal graph, and "with no lag" refers to the summary causal graph.

Type	Model	VAR(100, 5, 10)		VAR(50, 5, 10)		VAR(10, 3, 5)	
	Metric	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)	AUROC (↑)	AUPRC (↑)
with lag	cMLP	$0.973_{\pm 0.013}$	$0.770_{\pm 0.131}$	$0.995_{\pm 0.005}$	$0.893_{\pm 0.061}$	$0.994_{\pm 0.009}$	$0.912_{\pm 0.127}$
	JGC	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{0.998}_{\pm 0.003}$
	JRNGC-L1(ours)	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{0.998}_{\pm 0.004}$
	JRNGC-F(ours)	$\underline{0.997}_{\pm 0.001}$	$\underline{0.950}_{\pm 0.006}$	$\mathbf{1.000}_{\pm 0.001}$	$\underline{0.994}_{\pm 0.002}$	$0.999_{\pm 0.002}$	$\underline{0.993}_{\pm 0.013}$
with no lag	cMLP	$0.940_{\pm 0.0013}$	$0.851_{\pm 0.051}$	$0.973_{\pm 0.022}$	$0.929_{\pm 0.045}$	$0.978_{\pm 0.032}$	$0.923_{\pm 0.110}$
	cLSTM	$0.845_{\pm 0.045}$	$0.606_{\pm 0.102}$	$0.935_{\pm 0.018}$	$0.816_{\pm 0.051}$	$0.931_{\pm 0.061}$	$0.803_{\pm 0.162}$
	NAVAR(MLP)	$0.887_{\pm 0.023}$	$0.776_{\pm 0.040}$	$0.960_{\pm 0.019}$	$0.909_{\pm 0.043}$	$0.976_{\pm 0.131}$	$0.936_{\pm 0.086}$
	NAVAR(LSTM)	$0.860_{\pm 0.011}$	$0.750_{\pm 0.018}$	$0.930_{\pm 0.016}$	$0.851_{\pm 0.022}$	$0.952_{\pm 0.064}$	$0.875_{\pm 0.156}$
	JGC	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.001}$	$\underline{0.998}_{\pm 0.003}$
	CR-VAE	$0.645_{\pm 0.014}$	$0.268_{\pm 0.008}$	$0.638_{\pm 0.014}$	$0.282_{\pm 0.017}$	$0.749_{\pm 0.013}$	$0.401_{\pm 0.107}$
	JRNGC-L1(ours)	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.001}$	$\mathbf{0.999}_{\pm 0.003}$
	JRNGC-F(ours)	$\underline{0.984}_{\pm 0.002}$	$\underline{0.960}_{\pm 0.005}$	$\underline{0.998}_{\pm 0.001}$	$\underline{0.995}_{\pm 0.001}$	$0.996_{\pm 0.007}$	$0.994_{\pm 0.011}$

$D \in \{10, 50, 100\}$, maximum estimated lag $\eta \in \{5, 10\}$ and $\tau \sim N(0, I)$ to demonstrate the ability to deal with high-dimensional time series. We conduct experiments on five independently generated realizations of the system. From Table 1, we observe that using an input-output Jacobian matrix in the post-hoc analysis yields a more comprehensive and informative grasp of the Granger causal relationships within the data when compared to other analogous techniques like cMLP, cLSTM, NAVAR(MLP), NAVAR(LSTM), and CR-VAE. Although JGC also benefits from using the input-output Jacobian matrix to estimate Granger causality, our methods JRNGC-F and JRNGC-L1 offer distinct advantages. They effectively integrate the Jacobian matrix into both training and post-hoc phases, resulting in remarkable performance, especially when dealing with more complex datasets than the VAR dataset, such as the Lorenz dataset.

Lorenz-96. For this dataset, we simulated up to $T = 500$ observations and we set $D = 10, F \in \{10, 40\}$. F is a forcing constant, and higher values of F can lead to increased chaotic behavior, enhanced nonlinearity (Karimi & Paul, 2010a). When F is 40, the time series becomes more challenging to predict than when F is 10. We conduct experiments on five independently generated realizations of the system. As can be seen from Table 2, our methods JRNGC-F and JRNGC-L1 achieve the best scores and robust AUPRC scores, especially when the $F = 40$, i.e., the relationships between variables are more complex.

Both the VAR datasets and Lorenz-96 dataset possess true summary Granger causality and full-time Granger causality. From Table 1 and 2, we observe our methods consistently exhibit robust performance in estimating the full-time Granger causality (denoted as the "with lag" performance) and the summary Granger causality (denoted as the "with no lag" performance). In contrast, other methods either fail to

capture full-time causal relationships (e.g., the cLSTM) or experience evident performance degradation when attempting to learn full-time causal relationships (e.g., cMLP, JGC for Lorenz-96 dataset with F = 40). The datasets introduced below, namely fRMI, DREAM-3 datasets and CausalTime, exclusively feature true summary causal graphs.

fMRI data. Here, we select data from the first subject in the third set of simulations for the experiment. We ran five independent experiments on the selected subject. Table 3 summarizes the performance of each method on fMRI data. Importantly, the true variable usage rate in this experiment is only 14.67%, emphasizing the sparsity of the data. In such data, identifying positive connections between different brain regions becomes crucial. Hence, we need to pay close attention to the AUPRC score, which reflects our method's ability to capture these sparse relationships. Remarkably, our methods demonstrate the capability to achieve a competitive AUROC score while simultaneously outperforming in terms of AUPRC, making them well-suited for this challenging context.

DREAM-3 dataset. We followed the settings in the baseline method eSRU, and the experimental results for the algorithms marked with an asterisk (∗) were directly referenced from the SRU method (Khanna & Tan, 2020). From Table 4, we can observe that all the methods can not achieve good performance on the DREAM-3 dataset. This is because there are too few observations in the DREAM3 dataset, which makes the neural network method overfit the data. However, an important observation can be made regarding the minimal usage of true variables, highlighting the sparsity of the causality existed in the dataset. Subsequently, when referring to the results in Table 10 in the appendix, we can observe that our method not only achieves a comparable AUROC score but also maintains a competitive AUPRC

Table 2. Comparisons of AUROC and AUPRC for Granger causality among different approaches on Lorenz-96 dataset. The mean results with standard error are reported by averaging over 5 runs. Note that "with lag" refers to the full-time causal graph, and "with no lag" refers to the summary causal graph.

Type	Model	F=10		F=40	
	Metric	AUROC ($\uparrow$)	AUPRC ($\uparrow$)	AUROC ($\uparrow$)	AUPRC ($\uparrow$)
with lag	cMLP	$\underline{0.999}_{\pm 0.002}$	$\underline{0.992}_{\pm 0.016}$	$0.998_{\pm 0.001}$	$0.968_{\pm 0.016}$
	JGC	$0.998_{\pm 0.001}$	$0.979_{\pm 0.011}$	$\underline{0.992}_{\pm 0.006}$	$0.939_{\pm 0.020}$
	JRNGC-L1(ours)	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{0.999}_{\pm 0.000}$	$\mathbf{0.994}_{\pm 0.003}$
	JRNGC-F(ours)	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{0.999}_{\pm 0.001}$	$\underline{0.989}_{\pm 0.009}$
with no lag	cMLP	$\underline{0.997}_{\pm 0.006}$	$\underline{0.998}_{\pm 0.005}$	$0.984_{\pm 0.007}$	$0.968_{\pm 0.016}$
	cLSTM	$0.974_{\pm 0.028}$	$0.949_{\pm 0.068}$	$0.896_{\pm 0.007}$	$0.854_{\pm 0.008}$
	NAVAR(MLP)	$0.993_{\pm 0.004}$	$0.986_{\pm 0.008}$	$0.900_{\pm 0.021}$	$0.828_{\pm 0.052}$
	NAVAR(LSTM)	$0.993_{\pm 0.006}$	$0.988_{\pm 0.011}$	$0.891_{\pm 0.036}$	$0.810_{\pm 0.056}$
	JGC	$0.989_{\pm 0.006}$	$0.982_{\pm 0.010}$	$0.956_{\pm 0.024}$	$0.943_{\pm 0.016}$
	CR-VAE	$0.898_{\pm 0.020}$	$0.776_{\pm 0.041}$	$0.654_{\pm 0.030}$	$0.493_{\pm 0.023}$
	JRNGC-L1(ours)	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{0.995}_{\pm 0.002}$	$\mathbf{0.994}_{\pm 0.003}$
	JRNGC-F(ours)	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$	$\underline{0.992}_{\pm 0.006}$	$\underline{0.989}_{\pm 0.009}$

Table 3. Comparative performance of the summary causal graph for the first subject in fMRI. The mean results with standard error are reported by averaging over 5 runs.

Model	AUROC ($\uparrow$)	AUPRC ($\uparrow$)
cMLP	$0.875_{\pm 0.014}$	$\underline{0.726}_{\pm 0.013}$
cLSTM	$0.845_{\pm 0.029}$	$0.644_{\pm 0.079}$
NAVAR(MLP)	$0.840_{\pm 0.004}$	$0.477_{\pm 0.014}$
NAVAR(LSTM)	$0.742_{\pm 0.015}$	$0.320_{\pm 0.020}$
eSRU	$0.743_{\pm 0.047}$	$0.514_{\pm 0.057}$
CR-VAE	$0.516_{\pm 0.029}$	$0.152_{\pm 0.007}$
JGC	$\mathbf{0.912}_{\pm 0.007}$	$0.536_{\pm 0.009}$
JRNGC-L1(ours)	$0.845_{\pm 0.011}$	$0.707_{\pm 0.004}$
JRNGC-F(ours)	$\underline{0.898}_{\pm 0.001}$	$\mathbf{0.749}_{\pm 0.003}$

Table 4. AUROC scores of summary causal graph for the five datasets in DREAM-3, alongside corresponding true variable usage percentages. Here, E1 refers to E.coli-1, Y1 refers to Yeast-1, and so on.

Model	E1	E2	Y1	Y2	Y3
cMLP*	0.644	0.568	0.585	0.506	0.528
cLSTM*	0.629	0.609	0.579	0.519	0.555
TCDF*	0.614	0.647	0.581	0.556	$\underline{0.557}$
SRU*	0.657	$\underline{0.666}$	0.617	0.575	0.550
eSRU*	$\underline{0.660}$	0.629	0.627	0.557	0.550
NAVAR(MLP)	0.557	0.577	$\mathbf{0.652}$	$\underline{0.573}$	0.548
NAVAR(LSTM)	0.544	0.473	0.497	0.477	0.466
JGC	0.504	0.527	0.604	0.553	0.521
CR-VAE	0.502	0.494	0.525	0.518	0.501
JRNGC-F(ours)	$\mathbf{0.666}$	$\mathbf{0.678}$	$\underline{0.650}$	$\mathbf{0.597}$	$\mathbf{0.560}$
True variable usage	1.25%	1.19%	1.66%	3.89%	5.51%

score. This consistency across both metrics demonstrates the robustness of our approach in accurately identifying the sparse Granger causality relationships within the dataset.

CausalTime. For the new benchmark datasets, we followed the experimental settings described by (Cheng et al., 2024b). Aside from our algorithm, **JRNGC**, all additional results presented in Table 5 were directly reproduced from the CausalTime (Cheng et al., 2024b), which benchmarked the performance of several recent and representative causal discovery methods. As illustrated in Table 5, among the methods evaluated, JRNGC-F, alongside CUTS+ and LCCM, consistently emerged as top performers, and most of the baselines do not get AUROC > 0.9. In contrast, our proposed algorithm, JRNGC-F, demonstrates novel SOTA performance on the AQI and Traffic datasets, as evidenced by its AUROC metrics, achieving scores of AUROC 0.9279 ± 0.0011 and AUROC 0.7294 ± 0.0046 respectively. In addition, JRNGC-F demonstrates its reliability and robustness across all datasets, consistently achieving metrics well above the baseline threshold of 0.5.

Model complexity. The motivation behind proposing a unified neural Granger causality learning framework arises from the need for a single model to handle forecasting. However, the current neural Granger causality framework requires duplicating models for each variable to learn Granger causality, which is not only inelegant but also comes with drawbacks, as mentioned earlier. Overcoming the challenge of learning Granger causality with multiple models could lead to a substantial reduction in model parameters to one divided by D (this could be achieved by utilizing the same neural network or a simpler yet more effective neural network.), where D denotes the dimensionality of the variables. As illustrated in Figure 3, our proposed method, JRNGC, surpasses existing popular neural Granger causality methods with significantly lower model complexity and higher accuracy. For additional analyses, please refer to the appendix.

Table 5. Comparative performance on CausalTime benchmark datasets. We highlight the best and the second best in bold and with underlining, respectively.

Methods	AUROC			AUPRC		
	AQI	Traffic	Medical	AQI	Traffic	Medical
GC	$0.4538_{\pm 0.0377}$	$0.4191_{\pm 0.0310}$	$0.5737_{\pm 0.0338}$	$0.6347_{\pm 0.0158}$	$0.2789_{\pm 0.0018}$	$0.4213_{\pm 0.0281}$
SVAR	$0.6225_{\pm 0.0406}$	$0.6329_{\pm 0.0047}$	$0.7130_{\pm 0.0188}$	$0.7903_{\pm 0.0175}$	$0.5845_{\pm 0.0021}$	$0.6774_{\pm 0.0358}$
N.NTS	$0.5729_{\pm 0.0229}$	$\underline{0.6329}_{\pm 0.0335}$	$0.5019_{\pm 0.0682}$	$0.7100_{\pm 0.0228}$	$0.5770_{\pm 0.0542}$	$0.4567_{\pm 0.0162}$
PCMCI	$0.5272_{\pm 0.0744}$	$0.5422_{\pm 0.0737}$	$0.6991_{\pm 0.0111}$	$0.6734_{\pm 0.0372}$	$0.3474_{\pm 0.0581}$	$0.5082_{\pm 0.0177}$
Rhino	$0.6700_{\pm 0.0983}$	$0.6274_{\pm 0.0185}$	$0.6520_{\pm 0.0212}$	$0.7593_{\pm 0.0755}$	$0.3772_{\pm 0.0093}$	$0.4897_{\pm 0.0321}$
CUTS	$0.6013_{\pm 0.0038}$	$0.6238_{\pm 0.0179}$	$0.3739_{\pm 0.0297}$	$0.5096_{\pm 0.0362}$	$0.1525_{\pm 0.0226}$	$0.1537_{\pm 0.0039}$
CUTS+	$\underline{0.8928}_{\pm 0.0213}$	$0.6175_{\pm 0.0752}$	$\mathbf{0.8202}_{\pm 0.0173}$	$\underline{0.7983}_{\pm 0.0875}$	$\mathbf{0.6367}_{\pm 0.1197}$	$0.5481_{\pm 0.1349}$
NGC	$0.7172_{\pm 0.0076}$	$0.6032_{\pm 0.0056}$	$0.5744_{\pm 0.0096}$	$0.7177_{\pm 0.0069}$	$0.3583_{\pm 0.0495}$	$0.4637_{\pm 0.0121}$
NGM	$0.6728_{\pm 0.0164}$	$0.4660_{\pm 0.0144}$	$0.5551_{\pm 0.0154}$	$0.4786_{\pm 0.0196}$	$0.2826_{\pm 0.0098}$	$0.4697_{\pm 0.0166}$
LCCM	$0.8565_{\pm 0.0653}$	$0.5545_{\pm 0.0254}$	$\underline{0.8013}_{\pm 0.0218}$	$\mathbf{0.9260}_{\pm 0.0246}$	$0.5907_{\pm 0.0475}$	$\mathbf{0.7554}_{\pm 0.0235}$
eSRU	$0.8229_{\pm 0.0317}$	$0.5987_{\pm 0.0192}$	$0.7559_{\pm 0.0365}$	$0.7223_{\pm 0.0317}$	$0.4886_{\pm 0.0338}$	$\underline{0.7352}_{\pm 0.0600}$
SCGL	$0.4915_{\pm 0.0476}$	$0.5927_{\pm 0.0553}$	$0.5019_{\pm 0.0224}$	$0.3584_{\pm 0.0281}$	$0.4544_{\pm 0.0315}$	$0.4833_{\pm 0.0185}$
TCDF	$0.4148_{\pm 0.0207}$	$0.5029_{\pm 0.0041}$	$0.6329_{\pm 0.0384}$	$0.6527_{\pm 0.0087}$	$0.3637_{\pm 0.0048}$	$0.5544_{\pm 0.0313}$
JRNGC-F (ours)	$\mathbf{0.9279}_{\pm 0.0011}$	$\mathbf{0.7294}_{\pm 0.0046}$	$0.7540_{\pm 0.0040}$	$0.7828_{\pm 0.0020}$	$\underline{0.5940}_{\pm 0.0067}$	$0.7261_{\pm 0.0016}$

Table 6. Evaluation of the impact of Residual MLP Layer Depth and Jacobian regularizer. JR-F denotes F-norm Jacobian regularizer.

JR-F	Residual Layer	VAR100		fMRI		Lorenz-96	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
×	0	$0.969_{\pm 0.003}$	$0.667_{\pm 0.023}$	$0.764_{\pm 0.000}$	$0.266_{\pm 0.000}$	$0.813_{\pm 0.018}$	$0.571_{\pm 0.033}$
✓	0	$\mathbf{0.997}_{\pm 0.001}$	$\mathbf{0.950}_{\pm 0.006}$	$\mathbf{0.898}_{\pm 0.001}$	$\mathbf{0.749}_{\pm 0.003}$	$0.875_{\pm 0.010}$	$0.692_{\pm 0.036}$
×	1	$0.936_{\pm 0.009}$	$0.501_{\pm 0.032}$	$0.729_{\pm 0.027}$	$0.175_{\pm 0.023}$	$0.954_{\pm 0.021}$	$0.772_{\pm 0.061}$
✓	1	$0.993_{\pm 0.002}$	$0.894_{\pm 0.014}$	$0.836_{\pm 0.018}$	$0.448_{\pm 0.013}$	$0.995_{\pm 0.003}$	$0.959_{\pm 0.021}$
×	5	$0.957_{\pm 0.005}$	$0.606_{\pm 0.032}$	$0.766_{\pm 0.026}$	$0.227_{\pm 0.015}$	$0.995_{\pm 0.004}$	$0.959_{\pm 0.025}$
✓	5	$0.966_{\pm 0.008}$	$0.688_{\pm 0.047}$	$0.748_{\pm 0.032}$	$0.329_{\pm 0.013}$	$\mathbf{0.999}_{\pm 0.001}$	$\mathbf{0.989}_{\pm 0.009}$

4.3. Ablation studies

In this section, we conduct ablation studies to investigate the individual contributions of various components of our methodology to the overall performance. These experiments are designed to identify the critical factors that drive the efficacy of our residual MLP model equipped with Jacobian regularization. By selectively disabling certain features, we aim to provide insights into the essential elements of our approach and their impact on learning outcomes. This helps in understanding the robustness and sensitivity of our model under different configurations.

In the first experiment, we conducted model ablations using three datasets: VAR100, fMRI, and Lorenz-96. We evaluated the impact of omitting the input-output Jacobian regularizer and varying the number of residual network layers on the performance of the algorithm. The results are detailed in Table 6. The Jacobian matrix regularizer proved highly beneficial for capturing Granger causality. The advantage of adding residual layers was marginal for the linear VAR100 dataset but particularly advantageous for the nonlinear Lorenz-96 dataset. These findings confirm the versatility and efficacy of our approach across different scenarios and underscore the importance of tailoring the
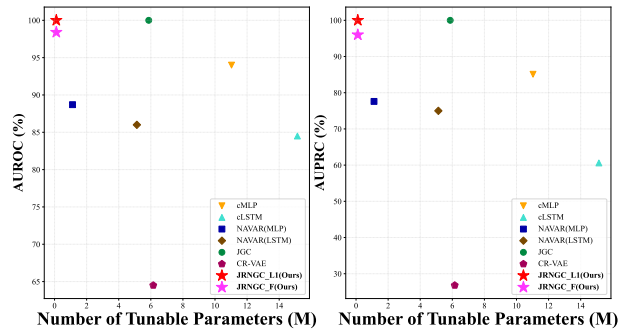

Figure 3. Performance comparisons on a 100-dimensional VAR dataset: AUROC, AUPRC, and the number of tunable parameters.

model to the specific characteristics of each dataset.

In the second experiment, we compare the impact of employing sparsity constraints versus Directed Acyclic Graph (DAG) constraints on the input-output Jacobian matrix using the VAR(100, 5, 10) and DREAM4 dataset (Marbach et al., 2009). Table 7 indicates that our approach is robust to the choice of ground-truth structure and performs better than

Table 7. Ablation study comparing the efficacy of different regularizers and ground-truth graph structures in detecting summary causal graphs. JR-DAG denotes the DAG regularizer, and JR-F denotes the F-norm regularizer, both applied to the input-output Jacobian matrix.

Dataset	Ground-truth Structure	Regularizer	AUROC	AUPRC
VAR (100, 5, 10)	DAG	JR-DAG	0.968	0.922
VAR (100, 5, 10)	DAG	JR-F	0.984	0.959
DREAM4	have-cycles	JR-DAG	0.614	0.042
DREAM4	have-cycles	JR-F	0.766	0.253

Table 8. AUROC and AUPRC performances on the Lorenz dataset ($F = 40$) when adopting the Jacobian regularizer to the LSTM.

Type	Model	AUROC ($\uparrow$)	AUPRC ($\uparrow$)
with lag	LSTM	0.956 ± 0.005	0.538 ± 0.045
	+JacoR	$\mathbf{0.992} \pm 0.003$	$\mathbf{0.929} \pm 0.014$
with no lag	LSTM	0.661 ± 0.040	0.539 ± 0.045
	+JacoR	$\mathbf{0.942} \pm 0.019$	$\mathbf{0.929} \pm 0.014$
	cLSTM	0.896 ± 0.007	0.854 ± 0.008

Table 9. Overfitting Analysis of JRNGC-F and same model without Jacobian regularizer (w/o JR) on Lorenz-96 dataset with $F = 40$.

Model	MSE	training loss	testing loss
JRNGC	0.878	0.0197	0.0830
w/o JR	1.269	0.0047	0.1088

the DAG regularizer on the input-output Jacobian matrix.

4.4. Alternative neural architecture

Our Jacobian regularizer functions independently of the network structure. We chose the residual-MLP architecture for its simplicity and efficiency. To show the regularizer's structure-agnostic nature, we conducted two experiments using the Lorenz-96 dataset with a forcing parameter of $F = 40$. The first experiment assessed the detection of summary causal graphs with a cLSTM model, a standard LSTM, and an LSTM enhanced by our Jacobian regularizer. Notably, the cLSTM model only identified the summary causal graph. The second experiment evaluated full-time causal relationships using both the standard LSTM and the enhanced LSTM. As shown in Table 8, the results confirm that our Jacobian regularizer is independent of the network's structure and effectively improves the model's ability to learn both summary and full-time Granger causality. More details can be seen in the appendix.

4.5. Causal discovery and time series forecasting

As the definition of Granger causality, its assessment is based on the performance of prediction. A better prediction model and a better Granger causality constraint will help each other. In other words, we believe that the Granger causality constraint will help the model avoid overfitting the data. We take an example of the Lorenz-96 dataset with $F = 40$ and see if the predictive performance will be enhanced by introducing the Granger causality constraint, i.e., the input-output Jacobian regularizer. Specifically, we compare the training and testing losses with and without Jacobian constraints. We also evaluate the mean squared error (MSE) between the predicted value and true value under both constrained and unconstrained conditions. The results depicted in Table 9 highlight how Granger causality contributes to enhancing the generalization capability of the predictive model and the causal constraint (i.e., input-output Jacobian regularizer) helps reduce overfitting and improve prediction performance.

5. Conclusion

In this work, we introduce a novel neural Granger causality method, termed Jacobian Regularizer-based Neural Granger Causality (JRNGC), which incorporates a new framework and an input-output Jacobian regularizer. By representing Granger causality through the input-output Jacobian matrix, our method eliminates the need to build separate models for each target variable and the sparsity on the weights of the first layer. Specifically, we utilize a joint model and apply a sparsity penalty to the input-output Jacobian matrix. Extensive experiments demonstrate that our method enhances the modeling of interactions within time series data and enables a more accurate assessment of both full-time Granger causality and summary Granger causality. Our code is available at https://github.com/ElleZWQ/JRNGC.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

References

Amblard, P.-O., Vincent, R., Michel, O. J., and Richard, C. Kernelizing geweke's measures of granger causality. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2012.

Assaad, C. K., Devijver, E., and Gaussier, E. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Bellot, A., Branson, K., and van der Schaar, M. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SsHBkfeRF9L.

Brouwer, E. D., Arany, A., Simm, J., and Moreau, Y. Latent convergent cross mapping. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=4TSiOTkKe5P.

Bussmann, B., Nys, J., and Latré, S. Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pp. 446–460. Springer, 2021.

Cheng, Y., Yang, R., Xiao, T., Li, Z., Suo, J., He, K., and Dai, Q. CUTS: Neural causal discovery from irregular time-series data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UG8bQcD3Emv.

Cheng, Y., Li, L., Xiao, T., Li, Z., Suo, J., He, K., and Dai, Q. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024a.

Cheng, Y., Wang, Z., Xiao, T., Zhong, Q., Suo, J., and He, K. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=iad1yyyGme.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., and Yu, R. Long-term forecasting with tide: Time-series dense encoder, 2023.

Entner, D. and Hoyer, P. O. On causal discovery from time series data using fci. *Probabilistic graphical models*, pp. 121–128, 2010.

Gong, W., Jennings, J., Zhang, C., and Pawlowski, N. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=i_1rbq8yFWC.

Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Hoffman, J., Roberts, D. A., and Yaida, S. Robust learning with jacobian regularization, 2019.

Jakubovitz, D. and Giryes, R. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 514–529, 2018.

Karimi, A. and Paul, M. R. Extensive chaos in the lorenz-96 model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(4), nov 2010a. doi: 10.1063/1.3496397. URL https://doi.org/10.1063%2F1.3496397.

Karimi, A. and Paul, M. R. Extensive chaos in the lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4):043105, 2010b.

Khanna, S. and Tan, V. Y. F. Economy statistical recurrent units for inferring nonlinear granger causality, 2020.

Li, H., Yu, S., and Principe, J. Causal recurrent variational autoencoder for medical time series generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8562–8570, Jun. 2023. doi: 10.1609/aaai.v37i7.26031. URL https://ojs.aaai.org/index.php/AAAI/article/view/26031.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239, 2009.

Marinazzo, D., Pellicoro, M., and Stramaglia, S. Kernel method for nonlinear granger causality. *Physical Review Letters*, 100(14), April 2008. ISSN 1079-7114. doi: 10.1103/physrevlett.100.144103. URL http://dx.doi.org/10.1103/PhysRevLett.100.144103.

Nauta, M., Bucur, D., and Seifert, C. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5(2):e9202, 2010.

Rhodes, T. and Lee, D. Local disentanglement in variational auto-encoders using jacobian l_1 regularization. *Advances in Neural Information Processing Systems*, 34: 22708–22719, 2021.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

Schreiber, T. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Sun, J., Taylor, D., and Bollt, E. M. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015.

Sun, X., Schulte, O., Liu, G., and Poupart, P. Nts-notears: Learning non-parametric dbns with prior knowledge. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pp. 1942–1964. PMLR, April 2023.

Suryadi, S., Chew, L. Y., and Ong, Y.-S. Granger causality using jacobian in neural networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 33(2), 2023.

Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. B. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Wiener, N. The theory of prediction. *Modern mathematics for engineers*, 1956.

Xu, C., Huang, H., and Yoo, S. Scalable causal graph learning through a deep neural network. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1853–1862, 2019.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, Jun. 2023. doi: 10.1609/ aaai.v37i9.26317. URL https://ojs.aaai.org/ index.php/AAAI/article/view/26317.

Zhou, W., Yu, S., and Chen, B. Causality detection with matrix-based transfer entropy. *Information Sciences*, 613: 357–375, 2022.

A. Background

A.1. Post-hoc analysis

Post-hoc analysis is a commonly used phase in deep learning after training a neural network. For instance, scientists often utilize heat maps to determine which features are more effective for classification. Similarly, time series forecasting models are designed to predict future values but not to explicitly capture Granger causality. Therefore, a post-hoc phase is necessary for Granger causality analysis. In our approach, the post-hoc phase can be summarized as follows: after training the multivariate time series forecasting model with Jacobian regularizer, our objective is to explore Granger causality across various time lags. To achieve this, we conduct significant tests on the learned input-output Jacobian matrix. The significance of the elements in the input-output Jacobian matrix determines the presence or absence of Granger causal relationships between variables. Specifically, if the value of an element in the Jacobian matrix is greater than zero, it indicates a Granger causal relationship between the variable corresponding to the row and the variable corresponding to the column. Conversely, if the value is zero or less than zero, it suggests no Granger causal relationship between the variables. The post-hoc analysis allows us to identify and quantify the Granger causality between different variables, providing valuable insights into the underlying dynamics of the multivariate time series.

A.2. Causal graphs for time series

Full-time causal graph and summary causal graph are always used for representing the causal relationships for time series (see Figure 4). It is always difficult to find the full-time causal graph, especially for the methods that use Long Short-Term Memory (LSTM) network (Tank et al., 2021)
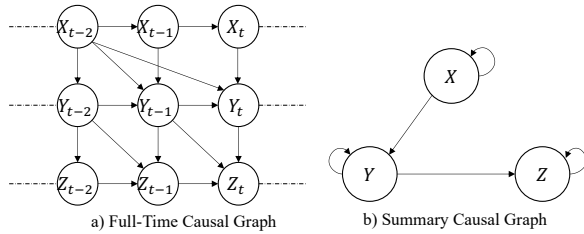

a) Full-Time Causal Graph   b) Summary Causal Graph

Figure 4. Typical causal graphs for time series. The full-time causal graph entails including all intricate causal links and interactions among variables. In contrast, the summary causal graph streamlines this complexity, focusing on the most significant causal connections.

A.3. AUPRC: an important metric for learning sparse causality

Sparse causality refers to situations where only a small subset of variables in a dataset actually exhibit causal relationships, while the majority remain unrelated. Variable usage percentage is a measure that quantifies the proportion of variables involved in causal relationships compared to the total number of variables. In our context, a low variable usage percentage signifies a sparse causal structure. For instance, if only 14.67% of variables are actively participating in causal relationships, it indicates a high degree of sparsity in the dataset. In the context of learning sparse causality, AUPRC is more informative than AUROC. A model with a high AUROC score could simply be good at identifying unrelated variable pairs while performing poorly in identifying the few relevant causal relationships among the sparse variables. On the other hand, a high AUPRC score indicates that the model is proficient in correctly identifying and prioritizing the sparse causal relationships. Therefore, AUPRC is a better indicator of the model's ability to capture and learn the sparse causal structure, making it a crucial metric for evaluating the performance of methods focused on sparse causality.

B. Datasets Details

VAR model. Vector autoregressions (VAR) model is defined as:

$$\mathbf{x}(t) = \sum_{\alpha=1}^{\tau} A_\alpha \mathbf{x}(t - \alpha) + \boldsymbol{\varepsilon}, \qquad (8)$$

where $\mathbf{x}$ is D-dimensional time series with time index t and τ is the maximum true time lag. A_α is the relation matrix between the D time series at time α. $\boldsymbol{\varepsilon}$ is a noise item. In our experiments on VAR datasets, we denote VAR(100, 5, 10) to represent a scenario where there are 100 dimensions in total. Among these, the true time lag τ is set to 5, while the maximum estimated lag η is set to 10.

Lorenz-96. It is a classic chaotic dynamical model, which describes the non-linear interaction between variables in systems such as atmospheric circulation and other natural phenomena (Karimi & Paul, 2010b), which is defined as:

$$\frac{dx_i(t)}{dt} = (x_{i+1}(t) - x_{i-2}(t)) x_{i-1}(t) - x_i(t) + F, \quad (9)$$

where t is the time index and the constant F is an important parameter representing the external forcing on the system. The sequence index i is taken modulo, i.e., $x_{-1} = x_{D-1}, x_0 = x_D$ and $i = 1, 2, 3, \cdots, D$.

fMRI data. It is usually used to estimate the brain network. Stephen M. Smith et al. generated rich, realistic simulated fMRI data for a wide range of underlying networks, experimental scenarios, and problematic confounders in the data

to compare different approaches to connectivity estimation (Smith et al., 2011). The dataset consists of 50 subjects, with each subject having 15 nodes and 200 observations.

DREAM-3 dataset. It is a realistic gene expression data set from the DREAM-3 challenge (Prill et al., 2010). This challenge includes five simulated datasets, comprising two E. coli datasets and three yeast datasets, each featuring a distinct underlying Granger causality plot. Each dataset contains 100 numbers of different time series, each with 46 replicates, sampled at 21 time points, resulting in a total of 966 time points. As can be seen, the data is very limited in length and is a difficult non-linear dataset.

DREAM-4 dataset. The DREAM-4 network inference challenge, established by (Marbach et al., 2009), aims to facilitate the learning of gene regulatory networks from gene expression data. The challenge comprises 5 independent datasets, each containing data from 10 different time-series recordings. These recordings capture the expression levels of 100 genes over 21 time steps.

CausalTime. CausalTime is proposed by (Cheng et al., 2024b) to evaluate time-series causal discovery algorithms in real applications.

C. Model complexity

While our method may not require the same level of attention as large-scale models in terms of parameter count, our contribution lies in successfully addressing the challenge of joint model learning in exploring Granger causal relationships among multiple time series without increasing the complexity of a single model. We could exploit the same model or a simpler but more effective model. As evidenced by our experimental results, our approach allows us to utilize a single model to investigate Granger causal relationships among multiple time series. Specifically, we analyze scenarios involving both 100-dimensional variables and 10-dimensional variables to compare our methods with other approaches. The superiority of our algorithm can be observed from various perspectives, including experimental time, model parameter count, AUPRC, and AUROC, as shown in Table 11 and 12.

D. Motivation

To articulate our motivation clearly, we employ a Q-A format to address a key inquiry.

Q: Why should we use the single neural network (NN) for Granger causality analysis?

A: The justification for employing a single NN model with shared hidden layers in the neural Granger causality framework is multifaceted. Firstly, this approach represents a pioneering effort in the field. To the best of our knowledge, it is the inaugural application of such a model configuration for multivariate Granger causality analysis. The innovative design of this model allows for an integrated processing of time series data, which is critical given the complex interdependencies among variables as illustrated in Figure 4. Conventional frameworks that apply Granger causality assumptions or regularizers directly to the weights of only the first or last layer tend to overemphasize the importance of these weights. This not only distorts the model's ability to accurately discern true causal relationships but also inadvertently magnifies the influence of non-causal variables in predictions. This misalignment can lead to incorrect conclusions about Granger causality, considering Granger causality's definition relies on enhancing prediction accuracy by including potentially causal variables' past values. Additionally, the traditional approach of employing multiple models to predict multiple variables is inefficient. It necessitates training separate models for each variable, which is both resource-intensive and time-consuming. By contrast, our single NN model streamlines this process, enhancing operational efficiency without compromising the quality of causality analysis. The performance metrics of our model, as demonstrated in our experimental results (see Table 11 and Table 12), attest to the effectiveness of using a unified framework. This approach not only simplifies the learning of causal relationships but also ensures a higher degree of accuracy in identifying and interpreting these relationships.

E. Implementation Experiment Details

E.1. Significance test

Considering the real scenes, we have no access to obtain the true causal relationships between variables. Instead, we aim to extract causal relationships from the estimated weighted causal matrix $\mathbf{W}$. However, the learned causal relationships will inherently include noise, i.e., when there is no causal relationship between variables i and j, $W_{i,j}$ or $W_{j,i}$ is challenging to be precisely equal to 0. Therefore, we need to test whether each element of the weighted causal matrix is significantly larger than 0 (Zhou et al., 2022; Suryadi et al., 2023).

Here, we use the significance test method as (Zhou et al., 2022), i.e., generating surrogate datasets for hypothesis testing. To generate surrogate datasets, we need to disrupt the intrinsic causal relationships within the data by eliminating the full-time interdependencies among variables. For instance, let's consider three variables: x_1, x_2, and x_3, each containing 500 observations. To create surrogate datasets without causal links between them, we form the initial dataset by concatenating $x_1[0:100]$, $x_2[m:100+m]$, and $x_3[2m:100+2m]$, with m set to a significantly large value. Specifically, we set m to 50, which surpasses the true

lag (where the true lag value is 3), and the resulting dataset has a time length of 100. With our 10-dimensional VAR dataset having 600 observations, we generate 50 sets of surrogate datasets. Training these surrogate datasets using the same model structure as the model for the original dataset, we derive the weighted causal matrix $\mathbf{Ws}$ for each surrogate dataset. As causal relationships are absent among variables in the surrogate datasets, we establish the threshold using the 3-sigma rule. Our threshold is defined as follows:

$$
\begin{aligned}
\mathrm{Ws}_\mu &= \frac{1}{\mathrm{Num}_s} \sum \mathbf{Ws}_i \\
\mathrm{Ws}_\sigma &= \sqrt{\frac{1}{\mathrm{Num}_s} \sum_{i=1}^{\mathrm{Num}_s} (\mathbf{Ws}_i - \mathrm{Ws}_\mu)^2}, \quad (10) \\
\mathrm{Thres} &= \mathrm{Ws}_\mu + 2\mathrm{Ws}_\sigma
\end{aligned}
$$

where $\mathbf{Ws}_i$ is the weighted causal matrix for the i-th surrogate dataset. Ws_μ is the mean of the weighted causal matrices. Ws_σ is the standard deviation of the weighted causal matrices, Num_s is the number of surrogate datasets. As shown in Figure 5 and 6, we have good performance in obtaining the summary and full-time causal graphs. Our method can estimate the causal graph of each time (e.g., $t - lag$) successfully.

E.2. More details on the experiment for DREAM-3 dataset

We supplement the experimental results of NAVAR(MLP), NAVAR(LSTM), JGC, CR-VAE, and our methods on the DREAM-3 dataset. Table 10 shows that our approach demonstrates superior performance compared to other algorithms.

E.3. More details for experiment for alternative neural architecture

In this experiment, the reason why the LSTM model can achieve full-time Granger causality is that we employed the same post-processing procedure as the one described in this paper. This approach was chosen to better highlight the effectiveness of our method.

E.4. Performance on a larger dataset

Although dream3 and fMRI are benchmark datasets, they have limitations for NN methods. We experimented with an epilepsy dataset, using 12 EEG time series selected from deeper brain structures with 4000 simulations. The dataset lacks causal ground truth. As shown in Fig 7, a), the result aligns with clinical practice, where doctors often operate in the last six areas.

Table 10. Comparisons of AUROC and AUPRC on DREAM-3 dataset.

Dataset	Model	AUROC	AURPC
E.colo-1	NAVAR(MLP)	0.557	0.102
	NAVAR(LSTM)	0.544	0.013
	JGC	0.504	0.018
	CR-VAE	0.502	0.013
	JRNGC-F	0.666	0.198
E.coli-2	NAVAR(MLP)	0.577	0.107
	NAVAR(LSTM)	0.473	0.012
	JGC	0.527	0.016
	CR-VAE	0.494	0.012
	JRNGC-F	0.678	0.202
Yeast-1	NAVAR(MLP)	0.652	0.073
	NAVAR(LSTM)	0.497	0.030
	JGC	0.604	0.026
	CR-VAE	0.525	0.017
	JRNGC-F	0.650	0.172
Yeast-2	NAVAR(MLP)	0.573	0.105
	NAVAR(LSTM)	0.477	0.038
	JGC	0.553	0.050
	CR-VAE	0.518	0.040
	JRNGC-F	0.597	0.142
Yeast-3	NAVAR(MLP)	0.548	0.089
	NAVAR(LSTM)	0.466	0.052
	JGC	0.521	0.059
	CR-VAE	0.501	0.055
	JRNGC-F	0.560	0.130

E.5. Jacobian regularizer hyperparameters sensitivity

Figure 7 b) illustrates the outcomes of varying the parameter λ. The findings indicate that our method exhibits robust performance across a range of values for both parameters. Additionally, with regards to the sensitivity of the number of random projections, it has been observed in (Hoffman et al., 2019) that once the simulation length surpasses 100, the number of random projections has minimal impact on the results.
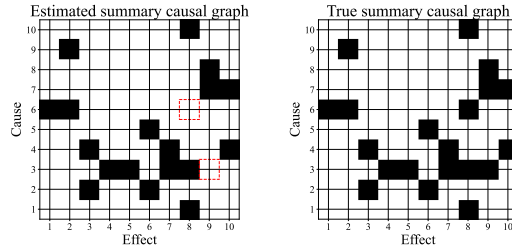
Figure 5. Estimated summary Causal graph Results on VAR (10,3,5) dataset. 95% confidence interval shown.
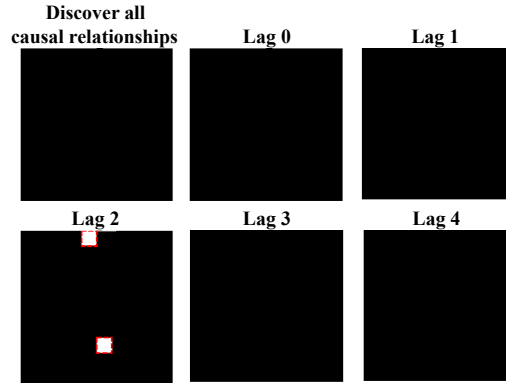
Figure 6. Performance of full-time causal graph learning. Black denotes correctly identified causal relationships, and white denotes overlooked or disregarded relationships. Lag indicates the time $t - \text{lag}$. Our method can detect full-time causal relationships. 95% confidence interval shown.

Figure 7. Granger Causality analysis by JRNGC on EEG dataset (left) and sensitivity analysis of λ (right) on Lorenz dataset.

Table 11. Comparison of tunable parameters and training time per epoch on the Var10 dataset.

Model	Number of tunable Parameters	Training time of an epoch(s)	AUROC (↑)	AUPRC (↑)
cMLP	52010	0.028	$0.978_{\pm 0.032}$	$0.923_{\pm 0.110}$
cLSTM	124510	0.031	$0.931_{\pm 0.061}$	$0.803_{\pm 0.162}$
NVAR(MLP)	<u>8110</u>	<u>0.017</u>	$0.976_{\pm 0.131}$	$0.936_{\pm 0.086}$
NVAR(LSTM)	111110	0.131	$0.952_{\pm 0.064}$	$0.875_{\pm 0.156}$
JGC	56600	0.203	$\mathbf{1.000}_{\pm 0.001}$	<u>0.998</u>$_{\pm 0.003}$
CR-VAE	390810	0.027	$0.749_{\pm 0.013}$	$0.401_{\pm 0.107}$
JRNGC-L1(ours)	**3110**	0.032	$\mathbf{1.000}_{\pm 0.001}$	$\mathbf{0.999}_{\pm 0.003}$
JRNGC-F(ours)	**3110**	**0.005**	<u>0.996</u>$_{\pm 0.007}$	$0.994_{\pm 0.011}$

Table 12. Comparison of tunable parameters and training time per epoch on the Var100 dataset.

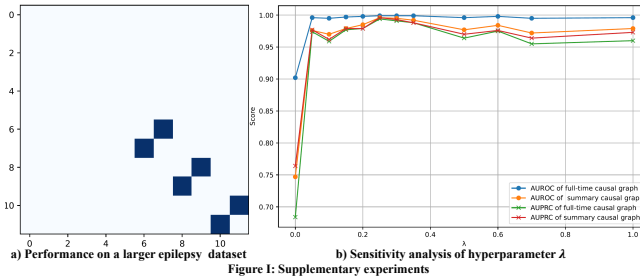| Model	Number of tunable Parameters	Training time of an epoch(s)	AUROC (↑)	AUPRC (↑)
cMLP	11030100	0.217	$0.940_{\pm 0.0013}$	$0.851_{\pm 0.051}$
cLSTM	15135100	0.433	$0.845_{\pm 0.045}$	$0.606_{\pm 0.102}$
NAVAR(MLP)	<u>1120100</u>	<u>0.175</u>	$0.887_{\pm 0.023}$	$0.776_{\pm 0.040}$
NAVAR(LSTM)	5130100	0.691	$0.860_{\pm 0.011}$	$0.750_{\pm 0.018}$
JGC	5870000	2.333	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$
CR-VAE	6150900	0.190	$0.645_{\pm 0.014}$	$0.268_{\pm 0.008}$
JRNGC-L1(ours)	**110300**	9.799	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{1.000}_{\pm 0.000}$
JRNGC-F(ours)	**110300**	**0.033**	<u>0.984</u>$_{\pm 0.002}$	<u>0.960</u>$_{\pm 0.005}$

F. Experimental Hyperparameters

Here, we present the tuned hyperparameters for our methods and the comparative approaches across various datasets in our experiments from Table 13 to Table 35.

Table 13. Tuned Hyperparameters of cMLP on the VAR Dataset.

	Hidden size	gate_lam	gate_regular	regular_lam	regular_type	Learning rate
Tuning range	[50,100]	[0.1,1]	['EL', 'GL', 'GSGL', 'H', 'SPH']	[0.01]	['lasso', 'ridge']	[5e-3,1e-1]
VAR $(10, 3, 5)$	[100]	0.1	'H'	0.01	'ridge'	5e-2
VAR $(50, 5, 10)$	[100]	0.01	'H'	0.01	'ridge'	5e-2
VAR $(100, 5, 10)$	[100,100]	0.001	'GL'	0.01	'ridge'	5e-2

Table 14. Tuned Hyperparameters of cMLP on the Lorenz-96 ($F = 10, 40$) Dataset.

	Hidden size	gate_lam	gate_regular	regular_lam	regular_type	Learning rate
Tuning range	[50,100]	[0.1,1]	['EL', 'GL', 'GSGL', 'H', 'SPH']	[0.01]	['lasso', 'ridge']	[5e-3,1e-1]
Lorenz-96 ($F = 10$)	[100]	0.1	'H'	0.01	'ridge'	5e-2
Lorenz-96 ($F = 40$)	[100]	0.1	'H'	0.01	'ridge'	1e-2

Table 15. Tuned Hyperparameters of cMLP on the fMRI Dataset.

	Hidden size	gate_lam	gate_regular	regular_lam	regular_type	Learning rate
Tuning range	[50,100]	[0.1,1]	['EL', 'GL', 'GSGL', 'H', 'SPH']	[0.01]	['lasso', 'ridge']	[5e-3,1e-1]
	[50]	0.1	'GL'	0.01	'ridge'	5e-2

Table 16. Tuned Hyperparameters of cLSTM on VAR Dataset.

	Hidden size	gate_lam	gate_regular	regular_lam	Learning rate
Tuning range	[50,100]	[0.1,1]	['EL', 'GL']	[0.01]	[5e-3,1e-1]
VAR $(10, 3, 5)$	50	0.1	'GL'	0.01	5e-2
VAR $(50, 5, 10)$	100	0.01	'GL'	0.01	5e-3
VAR $(50, 5, 10)$	100	0.01	'GL'	0.01	5e-3
VAR $(100, 5, 10)$	150	0.01	'GL'	0.01	5e-3

Table 18. Tuned Hyperparameters of cLSTM on the fMRI Datasets.

	Hidden size	gate_lam	gate_regular	regular_lam	Learning rate
Tuning range	[50,100]	[0.1,1]	['EL', 'GL']	[0.01]	[5e-3,1e-1]
	50	0.1	'GL'	0.01	5e-2

Table 19. Tuned Hyperparameters of NAVAR (MLP) on the VAR Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	Hidden Units	Layers	Batch Size	Learning rate	λ	μ
Tuning range	[5,10]	[50,100]	-	[32,128]	[5e-5,5e-3]	[0,1]	[1e-7,0.5]
VAR $(10, 3, 5)$	5	50	1	64	5e-4	0.02	9e-3
VAR $(50, 5, 10)$	10	100	1	64	5e-4	0.02	1e-4
VAR $(100, 5, 10)$	10	100	1	64	5e-4	0.02	1e-4

Table 20. Tuned Hyperparameters of NAVAR (MLP) on the Lorenz-96 ($F = 10, 40$) Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	Hidden Units	Layers	Batch Size	Learning rate	λ	μ
Tuning range	-	-	-	-	[5e-5,5e-3]	[0,1]	[1e-7,0.5]
Lorenz-96 ($F = 10$)	5	50	1	64	5e-4	0.5	9e-3
Lorenz-96 ($F = 40$)	5	50	1	64	5e-3	0.4	9e-3

Table 21. Tuned Hyperparameters of NAVAR (MLP, LSTM) on the fMRI Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

Model	K	Hidden Units	Layers	Batch Size	Learning rate	λ	μ
Tuning range	-	-	-	-	[5e-5,5e-3]	[0,1]	[1e-7,0.5]
NAVAR (MLP)	5	50	1	64	1e-4	0.5	9e-3
NAVAR (LSTM)	5	50	1	64	1e-4	0.2	8e-4

Table 22. Tuned Hyperparameters of NAVAR (MLP) on the DREAM-3 Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay. The tuned hyperparameters are copied from (Bussmann et al., 2021).

	K	Hidden Units	Layers	Batch Size	Learning rate	λ	μ
Tuning range	-	-	[1,4]	[16,256]	[5e-5,5e-3]	[0,0.5]	[1e-7,0.5]
Ecoli1	2	10	1	128	0.0005	0.1883	1.114e-4
Ecoli2	2	10	1	32	0.001	0.2011	1.710e-4
Yeast1	2	10	2	16	0.002	0.2697	1.424e-4
Yeast2	2	10	1	256	0.0002	0.1563	2.013e-4
Yeast3	2	10	1	16	0.0002	0.1559	1.644e-4

Table 23. Tuned Hyperparameters of NAVAR (LSTM) on the VAR Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	Hidden Units	Layers	Batch Size	Learning rate	λ	μ
Tuning range	[5,10]	[50,100]	-	[32,128]	[5e-5,5e-3]	[0,1]	[1e-7,0.5]
VAR $(10,3,5)$	5	50	1	128	1e-4	0.01	7.5e-4
VAR $(50,5,10)$	10	100	1	128	1e-4	0.01	7e-5
VAR $(100,5,10)$	10	100	1	128	1e-4	0.04	7e-5

Table 24. Tuned Hyperparameters of NAVAR (LSTM) on the Lorenz-96 ($F = 10, 40$) Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay.

	K	Hidden Units	Layers	Batch Size	Learning rate	λ	μ
Tuning range	-	-	-	-	[5e-5,5e-3]	[0,1]	[1e-7,0.5]
Lorenz-96 ($F = 10$)	5	50	1	64	5e-4	0.01	3e-3
Lorenz-96 ($F = 40$)	5	50	1	64	1e-4	0.078	8e-4

Table 25. Tuned Hyperparameters of NAVAR (LSTM) on the DREAM-3 Datasets. K is the Number of Lags Considered, λ is the Contribution Penalty, and μ is the Weight Decay. The tuned hyperparameters are copied from (Bussmann et al., 2021).

	K	Hidden Units	Layers	Batch Size	Learning rate	λ	μ
Tuning range	-	-	-	-	[5e-5,5e-3]	[0,0.5]	[1e-7,0.5]
Ecoli1	21	10	1	46	0.002	0.2208	1.094e-5
Ecoli2	21	10	1	46	0.002	0.1958	3.233e-6
Yeast1	21	10	1	46	0.002	0.2343	5.309e-5
Yeast2	21	10	1	46	0.002	0.2189	1.987e-5
Yeast3	21	10	1	46	0.002	0.2128	1.049e-5

Table 26. Hyperparameters of JGC on the VAR, Lorenz-96, fMRI, DREAM-3 . The hyperparameters are copied from the supplementary materials of the method JGC (Suryadi et al., 2023)

DATASET	MAX LAG	HIDDEN LAYERS	HIDDEN UNITS	TRAINING EPOCHS	LEARNING RATE	BATCH SIZE	SPARSITY HYPERPARAMS
VAR	[5 10]	2	50	2000	1e-3	64	$\lambda \in [0.5, 2.5]$
Lorenz96	5	2	50	2000	1e-3	64	$\lambda_{F=10} \in [0.5, 2.5] \lambda_{F=40} \in [0.5, 2.5]$
fMRI	1	2	50	2000	1e-3	64	$\lambda \in [1, 5]$
DREAM-3	10	2	50	2000	1e-3	64	$\lambda \in [0.5, 5]$

Table 27. Tuned Hyperparameters of CR-VAE. K is the Number of Lags Considered, λ is the sparsity-inducing penalty. λ-ridge is the ridge penalty at linear layer and hidden-hidden weights. Include_self is used for post-hoc analysis, i.e., generating the Granger weighted causal matrix.

	K	Hidden size	Learning rate	λ	λ-ridge	Include_self
Tuning range	[2,10]	[16,256]	[0.001,0.1]	[0.01,10]	[0,10]	[True False]
VAR $(10, 3, 5)$	5	100	0.1	0.1	0	False
VAR $(50, 5, 10)$	10	100	0.1	0.15	0	False
VAR $(100, 5, 10)$	10	100	0.05	0.1	0	False
Lorenz-96 $(F = 10)$	5	100	0.05	0.1	0	True
Lorenz-96 $(F = 40)$	5	100	0.1	0.05	0	True
fMRI	5	256	0.1	0.05	0	False
DREAM-3	2	100	0.1	0.005	0.1	True

Table 28. Tuned Hyperparameters of JRNGC-L1 on the VAR Datasets.

	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
VAR (10,3,5)	50	-	0	0.01	1e-3
VAR (50,5,10)	100	-	0	0.001	1e-3
VAR(100,5,10)	100	-	0	0.0001	1e-3

Table 29. Tuned Hyperparameters of JRNGC-L1 on the Lorenz-96 $(F = 10, 40)$ Datasets.

	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
Lorenz-96 $(F = 10)$	100	0.2	5	0.00001	1e-3
Lorenz-96 $(F = 40)$	100	0.2	5	0.02	1e-5

Table 30. Tuned Hyperparameters of JRNGC on the fMRI Datasets.

Model	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
JRNGC-L1	100	-	0	0.01	1e-3
JRNGC-F	100	-	0	1	2e-5

Table 31. Tuned Hyperparameters of JRNGC-F on the VAR Datasets.

	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
VAR (10,3,5)	50	-	0	0.01	1e-3
VAR (50,5,10)	50	0.2	5	0.0001	1e-3
VAR(100,5,10)	100	-	0	0.01	1e-3

Table 32. Tuned Hyperparameters of JRNGC-F on the Lorenz-96 ($F = 10, 40$) Datasets.

	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
Lorenz-96 ($F = 10$)	100	0.2	5	0.0001	1e-3
Lorenz-96 ($F = 40$)	100	0.2	5	0.2	1e-5

Table 33. Tuned Hyperparameters of JRNGC on the DREAM-3 Datasets.

Model	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
JRNGC-F	100	0.2	1	0.001	1e-3

Table 34. Tuned Hyperparameters of JRNGC on the DREAM-4 Datasets.

Model	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
JRNGC-F	100	0.2	0	1	1e-3

Table 35. Tuned Hyperparameters of JRNGC-F on the CausalTime Datasets.

	Hidden size	Dropout rate	Residual Layers	Jacobian lam	Learning rate
Tuning range	[50,100]	-	[0,5]	[1e-5,1]	[1e-5,1e-3]
AQI	50	0.2	0	0.002	2e-3
Traffic	50	0.2	5	0.001	1e-3
Medical	100	0.2	0	0.022	1e-3