Hans Journal of Data Mining 数据挖掘, 2025, 15(2), 184-200
Published Online April 2025 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2025.152016

Hans汉斯

# 动态正则格兰杰因果学习方法

姚牧芸[1]，王志海[1]，刘海洋[1*]，蒋文睿[1]，程帅卿[1]，蔡智明[2]，任　佳[2]，杨艳超[3]

[1]北京交通大学计算机科学与技术学院，北京
[2]中西创新学院数字科技学院，澳门
[3]中西创新学院国际语言服务研究院，澳门

收稿日期：2025年3月17日；录用日期：2025年4月9日；发布日期：2025年4月16日

## 摘　要

在医学和金融学等实际领域中，了解动态系统中的底层结构关系对于调节系统中的变量和预测系统未来状态至关重要。系统的动态变化会生成时间序列数据，通过观察时间序列数据可以分析系统的底层结构。格兰杰因果关系分析方法可以应用于一维或多维时间序列系统，现有的方法以组件式的建模方式分析每个系统变量特定的因果关系，受限于时间方向的强假设性和组件模型的单一性，其无法准确地挖掘出时间序列中的因果关系结构。本文提出了一种基于动态稀疏正则化的格兰杰因果发现方法DRGC (Dynamic Regularity Granger Causlity)。DRGC模型从卷积网络的输入权重中周期性地发掘变量在时间维度上的依赖信息，并以此为据向网络施加稀疏惩罚，以获得精确的格兰杰因果关系；同时，使用采样输入的循环网络提取数据中的长程依赖关系，同步优化卷积网络的权重，增强了模型发现因果关系的精确性和稳定性。在模拟数据集和真实系统生成的数据集上的实验表明，DRGC优于最先进的基线方法。

## 关键词

多维时间序列，格兰杰因果关系，深度神经网络，稀疏惩罚

# Dynamic Regularized Granger Causality Learning Method

**Muyun Yao[1], Zhihai Wang[1], Haiyang Liu[1*], Wenrui Jiang[1], Shuaiqing Cheng[1], Zhiming Cai[2], Jia Ren[2], Yanchao Yang[3]**

[1]School of Computer Science & Technology, Beijing Jiaotong University, Beijing
[2]Faculty of Digital Science and Technology, Macau Millennium College, Macau
[3]Institute of International Language Services Studies, Macau Millennium College, Macau

Received: Mar. 17th, 2025; accepted: Apr. 9th, 2025; published: Apr. 16th, 2025

*通讯作者。

文章引用: 姚牧芸, 王志海, 刘海洋, 蒋文睿, 程帅卿, 蔡智明, 任佳, 杨艳超. 动态正则格兰杰因果学习方法[J]. 数据挖掘, 2025, 15(2): 184-200. DOI: 10.12677/hjdm.2025.152016

## Abstract

**In practical fields such as medicine and finance, understanding the underlying structural relationships in dynamic systems is crucial for regulating system variables and predicting the system's future state. The dynamic changes of a system generate time series data, and by observing these time series, the underlying structure of the system can be analyzed. Granger causality analysis methods can be applied to univariate or multivariate time series systems. Existing methods analyze the specific causal relationships of each system variable using a modular modeling approach. However, these methods are limited by strong assumptions regarding the time direction and the simplicity of the modular models, which prevents them from accurately uncovering the causal relationship structure in multivariate systems. This paper proposes a Granger causality discovery method based on dynamic sparse regularization, called DRGC (Dynamic Regularized Granger Causality). The DRGC model periodically uncovers the temporal dependencies of variables from the input weights of a convolutional network and applies sparse penalties to the network accordingly, to obtain precise Granger causal relationships. Additionally, a cyclic network is used to extract long-range dependencies from the sampled input data, and the convolutional network's weights are optimized simultaneously, which enhances the precision and stability of the model in discovering causal relationships. Experiments conducted on simulated datasets and real-world system-generated datasets show that DRGC outperforms state-of-the-art baseline methods.**

## Keywords

**Multivariate Time Series, Granger Causality, Deep Neural Networks, Sparse Penalty**



## 1. 引言

从多维时间序列数据中挖掘变量间的因果关系、学习动态系统的内部结构，对时间序列的精准预测和干预分析等都有极大帮助。这种内部结构揭示了关于变量内部和变量之间的同期和滞后关系的信息。例如，在神经科学领域中，大脑活动通过各脑区域传播，不同区域的检查指标会随活动的传播而产生波动[1]，而定义这些指标的内部结构和活动的传播逻辑就是产生真实数据的原因。过去几十年，研究人员致力于从观测时间序列数据中发现因果关系[2]，取得了很大进展[3]-[6]。在诸多因果发现的方法中，格兰杰因果关系分析是最常用的一种框架[7] [8]，能够量化一个时间序列的过去值是否有助于预测另一个时间序列的未来演变趋势。

格兰杰因果关系取决于所研究的整个时间序列系统活动，故更适合分析高维的复杂序列数据。在多维的情况下，格兰杰因果分析可以捕获变量与其他变量过去值之间的因果关系并计算其大小。格兰杰因果关系分析有基于模型的方法和无模型的方法，大多数经典的基于模型方法都有“线性时间动力学”的假设，最具代表性的就是向量自回归模型[9]。在这种假设下，过去值对未来值的影响关系是线性的，线性系数的大小就是格兰杰因果效应的量化体现。使用向量自回归模型(vector autoregressive model, VAR)估计格兰杰因果关系需要设定一个最大考虑的时间滞后值，这种设定存在的问题是指定滞后值太大或太小都会影响格兰杰因果关系的评估效果。解决方法通常是通过添加 Lasso 或截断惩罚等策略来实现自动的滞后选择[4]，同时实现格兰杰因果关系的稀疏网络诱导，使得高维系统中每个变量只选择出少量与自己

有因果关系的变量。无模型的方法不再有基于模型方法的强线性相关性假设[10]，可以在变量之间发现非线性因果关系[1] [11]，而缺陷在于估计的结果方差较大，且当序列数量增加时，会受到维度灾难影响[2]，在较高维度的场景下难以适用。

早期适应时间序列的神经网络变体在多变量时间序列预测方面表现出令人满意的效果[12] [13]。长期以来，神经网络在多维时间序列系统中的应用还存在两个问题：第一，在变量较多的情况下进行联合建模会导致巨量的网络参数，大大增加了模型训练时的开销与复杂度；第二，联合建模的方法等价于将多维时间序列系统视为矢量时间序列，其本质仍然是一种黑箱算法，内部的结构对所预测系统的结构没有相当的解释作用。以上两个问题的存在使得联合建模多维时间序列不适用于格兰杰因果关系分析。

神经格兰杰因果方法[4]的提出为解决上述问题提供了优秀的思路。它为每个变量构建单独的神经网络，使用 MLPs 和 RNNs 实现了“组件级”的因果分析。这种建模方式在大大减少网络参数数量、加快模型训练速度的同时，通过解析网络参数使得模型有了可解释性。然而，得益于循环体系结构模拟长程依赖，组件级循环网络完全回避了时滞因果选择的问题，只能获取变量间整体上的因果强度，而无法挖掘出变量之间的因果关系在时间跨度上的差异。同时，采用 group Lasso 或分层式的稀疏惩罚方式给数据在时间方向上的因果关系大小施加了假设，即距离当前时刻较近的时滞值的因果影响更大，忽略了实际系统中可能因为采样率差异而导致的序列值影响不一定随着时间推移而减弱的问题。

为了解决上述问题，本文提出了动态正则格兰杰因果学习模型(Dynamic Regularized Granger Causality, DRGC)，通过采样因果图的方式打通循环网络与线性网络输入层之间的通道，使得循环网络发挥其捕捉长程依赖的能力来协助时滞选择，确保在线性网络难以拟合短时间序列时提供可靠的拟合性能。同时，本文设计了一种动态的网络输入权重分级惩罚策略，以增强线性网络在时滞上选择因果关系的正确性。本文的主要贡献如下：

1) 构建单维度建模的线性因果发现网络，并对提取出的因果关系进行采样，使用采样处理的时间序列数据输入循环网络进行训练，增强了线性网络在时间维度上挖掘因果关系的合理性，并且充分利用循环网络长程依赖的特性来协助因果选择。

2) 设计基于分级群组 Lasso 惩罚的动态稀疏惩罚策略，模型在训练过程中自动判断不同时滞之间的数值关系，协调不同原因变量各个时滞上因果关系的大小。

3) 在具有代表性的模拟数据集和拟真数据系统上验证提出模型的有效性，与当下先进的格兰杰因果发现方法对比均取得了性能的提升。

## 2. 相关工作

Granger 提出的格兰杰因果方法[7]，通过测试时间序列对预测另一个时间序列的帮助，被广泛应用于分析时间序列领域的因果关系。格兰杰因果分析最初假设线性模型和因果结构可以通过拟合向量自回归(Vector Autoregression, VAR)模型来发现。后来，格兰杰因果关系的概念被扩展到非线性情况[8]。由于与新兴的深度神经网络有着高度兼容性，格兰杰因果关系的研究可以被扩展到更多复杂数据的情况，分析更深层次或包括混杂在内的时间序列因果关系。

格兰杰因果分析可以分为基于模型的方法和无模型的方法，大多数基于模型的方法会假设线性相关，使用自回归模型[14]。它假设序列过去值对目标序列(被预测序列)的未来值具有线性效应，且非零系数可以量化格兰杰因果效应的大小。通过添加 Lasso [15]或 Group Lasso 这类稀疏诱导正则化方法有助于将自回归模型中的线性格兰杰因果扩展到高维环境下[14] [16]。线性相关的假设可能会导致对实际中非线性关系的误解，并可能因为简化程度过高而产生对内部结构的不一致估计。无模型的方法可以克服在线性假设下实际观测之间的非线性依赖关系，它对潜在关系的假设最少，包括转移熵[1]或有向信息[11]。无模型

方法估计的结果可能由于高自由度而具有高度不确定性，且对数据量的需求较大，不适用于维数大幅增多的情况[17]。

深度神经网络兴起并发展以来，研究人员尝试将神经网络应用于因果发现的方法中。DYNOTEARS方法[6]利用深度学习模型扩展了基于分数的方法来学习 SVAR 模型，也称为动态贝叶斯网络(Dynamic Bayesian Network, DBN)。随着节点数量增加，网络扩展效果非常好，但 DYNOTEARS 框架仍基于线性 VAR 模型。更多的研究人员利用神经网络来推断非线性格兰杰因果关系，克服了许多传统时间序列因果发现框架的缺点。Wu 等人[18]的研究介绍了一种新的最小预测信息正则化方法，从时间序列推断因果关系，允许深度学习模型发现非线性因果关系。Xu 等人[19]提出了一种基于深度神经网络的可扩展因果发现算法。Singh 等人[20]进行了个体成对的格兰杰因果检验研究。为了将非线性相互作用纳入格兰杰因果关系检测，Tank 等人[4]提出了一类非线性结构，包含多层感知器(MLP)和递归神经网络(RNN)。NTiCD算法[21]利用多种神经网络的能力组合，通过连续优化技术来捕获时间序列内部和之间不随时间变化的因果关系。CUTS 算法[22]使用神经网络设计了潜在因果发现和数据填补两个交替阶段，用于将非线性格兰杰因果分析应用到不规则时间序列数据上，它的后续版本 CUTS+ [23]通过引入 C2FD 技术和图神经网络提高了方法在不规则数据上的可扩展性。

在可解释性方面，神经格兰杰因果发现[4]使用多层感知机和递归神经网络，通过解析神经网络的参数来获得可解释的非线性格兰杰因果关系。获得可解释因果关系的关键在于单维度输出序列的独立模型。Horvath 等人[24]的研究提出了学习型核函数 LeKVAR 和一种解耦时滞和个体时间序列的机制，实现了有延迟的时滞选择和因果解释，提供了更好的伸缩性。

总体来说，当下神经格兰杰因果发现方法实现了减轻模型参数量和时滞选择，同时大大增强了模型的可解释性。然而目前的方法仍然存在模型在短时间序列下性能不佳和时滞发掘不准确等问题。

## 3. 时间序列及其因果关系模型

### 3.1. 多维时间序列与因果图

多维时间序列数据可以是由一个实际的系统中多个变量随时间进行采样观测得到，其在形状上可以理解为一个二维矩阵结构。从序列内部的相互关系角度出发，可以理解为多个单维时间序列的拼接，也可以理解为一个矢量随时间变化而产生的一系列值。多维时间序列的数学模型由式(1)给出

$$x_t = \left(x_{t1}, x_{t2}, \cdots, x_{tp}\right). \tag{1}$$

其中 $p$ 是系统中的变量个数，$t = 1, 2, \cdots, T$ 代表时间值。在数学模型中，由于没有考虑实际系统中的采样率因素，时间值上相邻的值仅代表数值的先后关系，而不包含时间维度上的距离信息。

在实际系统中，由于变量之间存在依赖，系统中各个变量会构成因果关系。因此可以为时间序列数据建立一个因果体系以指示变量之间相互影响的关系与程度大小，称为因果图。在数值上，因果图可以表示为一个 $p \times p$ 大小的方阵，$G \in \mathbb{R}^{p \times p}$。$G$ 中的每一项 $c_{ij}$ 代表变量 $j$ 对变量 $i$ 的影响程度，如图 1(a)所示。从时间维度上看，原因变量对其结果变量的影响可以细化到不同的时刻，通过在时间维度上对全局因果图进行扩展，可以得到时滞因果图，“时滞”指时间滞后值，即当前时刻之前某时刻的值。时滞因果图是一个三维图形，可以对不同结果变量进行划分，得到单一变量的时滞因果图，如图 1(b)所示。

### 3.2. 格兰杰因果关系模型

假设变量 $x_t \in \mathbb{R}^p$ 是一个 $p$ 维的平稳时间序列，并假设数据是在一定时间 $1 \le t \le T, t \in Z$ 内观测到的。时间序列的线性格兰杰因果关系一般使用 VAR 模型来进行研究。在 VAR 模型中，时间序列 $x$ 在时刻 $t$

处的值 $x_t$ 被认为是该序列在过去的 $K$ 个时间滞后值的线性组合，如式(2)所示。

$$x_t = \sum_{k=1}^{K} A_k x_{t-k} + e_t \tag{2}$$

其中 $K$ 是一个正整数，可以被解释为假设的最大时间滞后阶数，即当 $k > K$ 时，$x_{t-k}$ 在任何维度上不会影响 $x_t$。$A_k \in \mathbb{R}^{p\times p}$ 是对应 $k$ 滞后值的因果矩阵。$e_t \in \mathbb{R}^p$ 是当前时刻的噪声误差，可以服从特定的零均值噪声分布。通过对每一阶滞后的值 $x_{t-k}$ 经过 $A_k$ 变换得到这一阶滞后对当前时刻值的潜在影响。

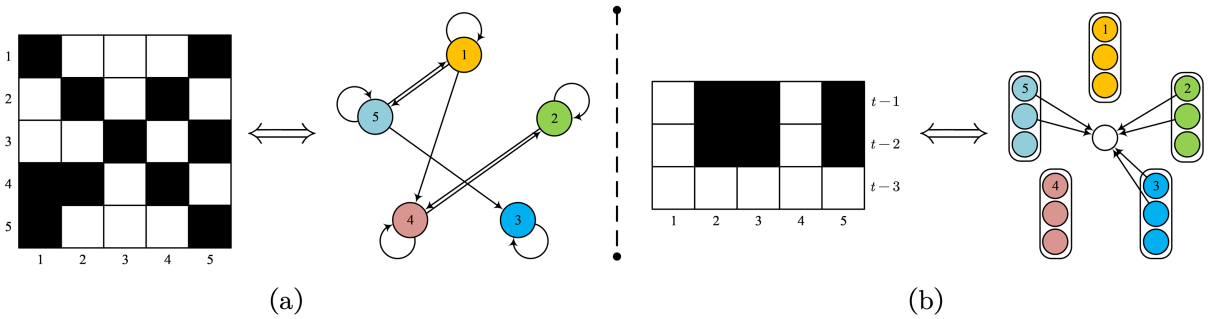


**Figure 1.** (a) Global causal graph; (b) Lag causal graph of univariate time series
**图 1.** (a) 时间序列数据的全局因果图；(b) 单变量时间序列时滞因果图

在类似金融系统等有高维度时间序列的领域中，数据之间往往不满足前面提到的线性关系，某变量过去的时间值可能会通过非线性函数作用于另一变量当前时刻的值。广义格兰杰因果关系见定义 1。

**定义 1**. 在多维时间序列 $x$ 中，若对所有的 $\left(x_{<t1},\cdots,x_{<tp}\right)$ 和 $x'_{<tj} \neq x_{<tj}$，有

$$g_i\left(x_{<t1},\cdots,x_{<tj},\cdots,x_{<tp}\right) = g_i\left(x_{<t1},\cdots,x'_{<tj},\cdots,x_{<tp}\right)$$

则称变量 $j$ 与变量 $i$ 之间不构成格兰杰因果关系，即 $g_i$ 对 $x_{<tj}$ 是不变的。若某一对 $i$ 和 $j$ 不满足上述条件，则称变量 $j$ 与变量 $i$ 之间构成格兰杰因果关系。其中 $j$ 为原因变量，$i$ 为结果变量。$x_{<tj}$ 代表序列 $j$ 在 $t$ 时刻之前的序列值。

## 4. 动态正则格兰杰因果学习模型

本文提出的模型总体结构如图 2 所示，由两个主要部分组成：线性因果发现网络和采样循环网络。线性网络以多层感知器的形式被构建，每个网络用于单独拟合时间序列数据中所有变量的 $K$ 阶滞后值到某变量当前时刻值的映射函数；循环网络以长短期记忆网络的形式被构建，每个变量对应的循环网络都需要进行独立的因果图采样并处理其输入的原始序列数据。

两部分模型的训练是交替进行的，在每一个训练轮中，使用线性网络作用于序列数据，得到线性输出预测误差项；将线性网络的输入权重按照滞后阶数进行提取并归一化，使用动态正则策略对权重进行稀疏约束，得到正则项；从线性网络的输入权重中提取出目标变量的单维度因果图，对原始序列数据进行覆盖处理，使用循环网络作用于处理后的序列数据上，得到循环输出预测误差项。通过各个误差项求和得到损失函数并进行优化，使模型逐步收敛至可以精确模拟因果作用机制。

### 4.1. 线性因果发现网络

本文在线性部分为每个变量单独建立网络模型，如式(3)所示

$$x_{ti} = g_i\left(x_{<t1},\cdots,x_{<tp}\right) + e_{ti} \tag{3}$$

其中，$x_{ti}$ 代表时间序列 $x_i$ 在 $t$ 时刻时的值，$e_{ti}$ 代表对应单维序列在 $t$ 时刻的随机噪声值，$g_i$ 是一个函数，指定如何将多维时间序列 $x$ 的 $t$ 时刻前的“过去值”映射到序列 $i$ 上。对多维序列中的每一个变量建立一个 MLP，然后将所有的网络并行组合在一起得到输出，如图 3(a)所示。

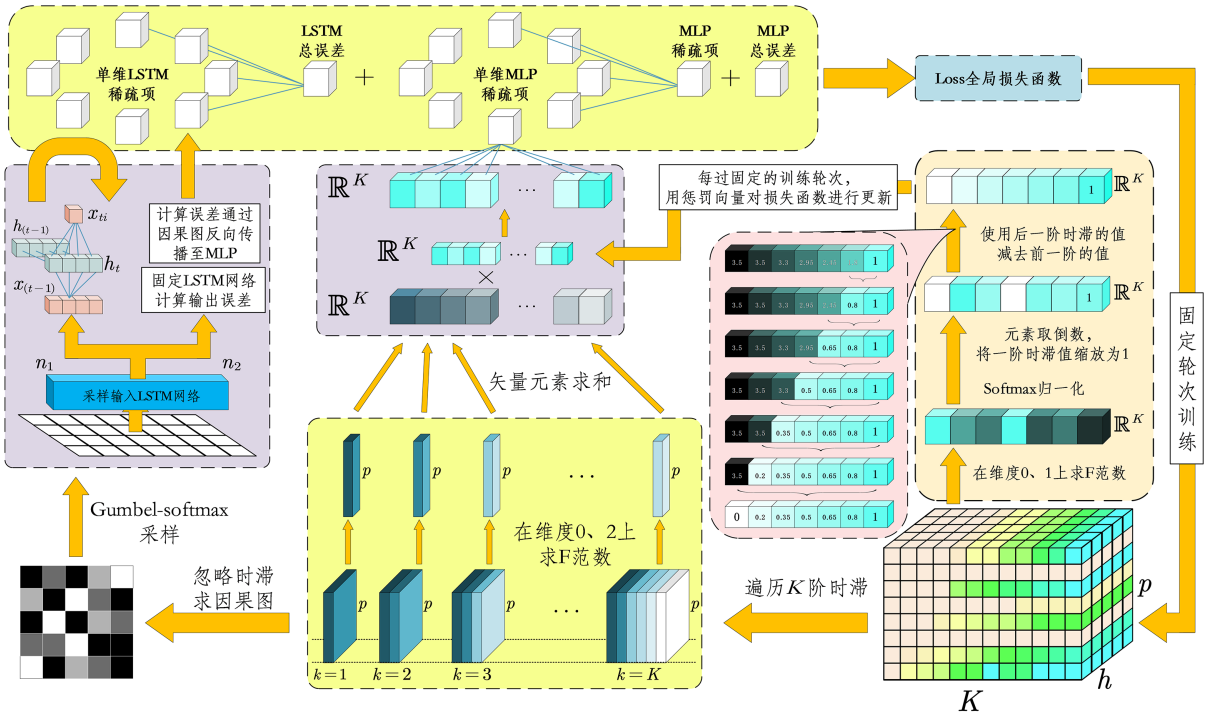


**Figure 2.** Overall model structure
**图 2.** 总体模型结构

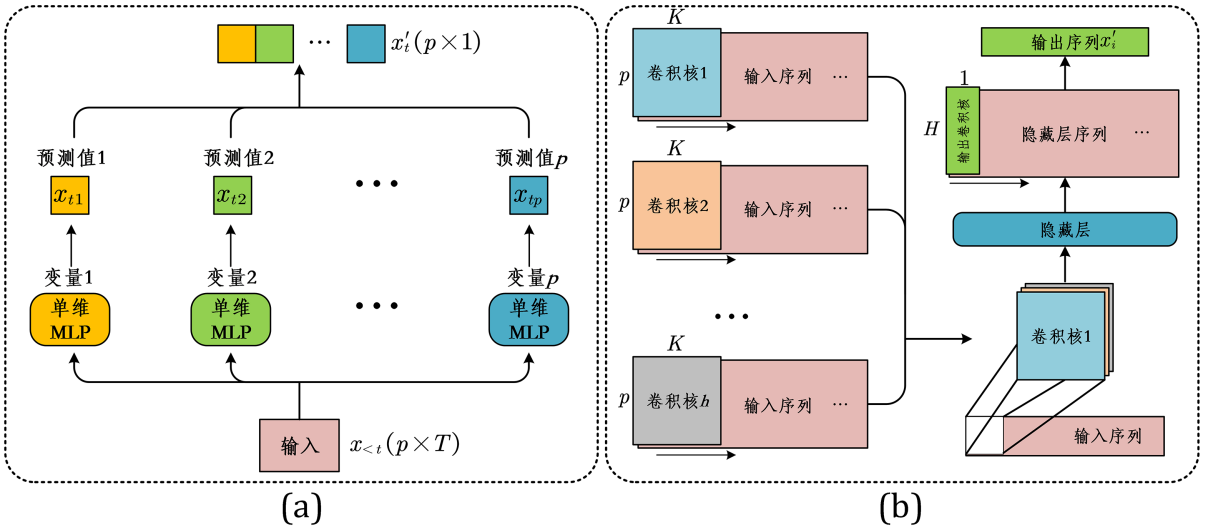


**Figure 3.** (a) Overall structure of component-level MLP; (b) Single-dimensional MLP structure
**图 3.** (a) 组件级 MLP 的总体结构；(b) 单维度 MLP 结构

在单维度线性网络的第一层，有 $H$ 个大小为 $p\times K$ 的一维卷积核，其中 $H$ 为可设置的隐藏单元维度，$p$ 为输入的多维时间序列的维度，$K$ 为预设定的最大可能时间滞后值。经过第一层卷积网络后得到维度为

$H$ 的隐藏层序列，将隐藏序列输入到后续的隐藏线性层中，最后经过一个卷积核大小为 $H\times1$ 的一维卷积层，得到网络的最终输出 $x_i'$，如图 3(b)所示。

本文以上述结构来模拟式(3)中的 $g_i(\cdot)$ 函数，在此结构中，网络的参数由每一层的权重 $\boldsymbol{W}$ 和偏置 $\boldsymbol{b}$ 确定，其中 $\boldsymbol{W}=\{W^1,\cdots,W^L\}$，$\boldsymbol{b}=\{b^1,\cdots,b^L\}$，$L$ 代表包括第一层线性层、隐藏线性层和最后一层线性层在内的总层数。cMLP 的可解释性主要体现在其第一层的网络权重中。第一层的权重结构是一个大小为 $H\times p\times K$ 的三维张量，即 $W^1\in\mathbb{R}^{H\times p\times K}$。对第一层权重进行分解，$W^1=\{W^{11},\cdots,W^{1K}\}$，$W^{1k}\in\mathbb{R}^{p\times H}$，$k=1,\cdots,K$。网络的其他参数维度为：$W^l\in\mathbb{R}^{H\times H}\ (l=2,\cdots,L-1)$，$W^L\in\mathbb{R}^H$，$b^l\in\mathbb{R}^H\ (l=1,\cdots,L-1)$，$b^L\in\mathbb{R}$。基于上述权重表达，输入数据在时间 $t$ 处经过输入层后的隐藏向量由式(4)给出

$$h_t^1=\sigma\left(\sum_{k=1}^{K}W^{1k}x_{t-k}+b^1\right) \tag{4}$$

其中，$\sigma$ 是激活函数，可以是 logistic 函数或 relu 函数，后续隐藏层中的隐藏向量使用 $h_t^l$ 表示，使用相同的激活函数 $\sigma$，$h_t^l=\sigma\left(W^lh_t^{l-1}+b^l\right)$。通过 $L-1$ 个隐藏层后，输出的单维度序列 $x_{ti}$ 由最后一个隐藏层所有单元的线性组合表示，$x_{ti}=W^Lh_t^{L-1}+b^L+e_{ti}$，其中误差项 $e_{ti}$ 由零均值高斯噪声分布建模。

## 4.2. 采样循环因果发现网络

循环网络能更好地提取时间维度上数据的变化趋势与不同时刻值之间的关系，因此当时间序列数据点有限时，循环网络呈现出比线性网络更好的因果图发现结果。

本文对 MLP 网络提取出的单维度因果图进行伯努利采样操作，用采样出的因果图对时间序列数据进行覆盖处理，使得输入 LSTM 网络的数据尽可能只保留对目标变量有因果影响的数据，消除其他变量的影响，如图 4(a)所示。得到处理过的数据之后，将其输入对应维度的 LSTM 网络进行计算，并用输出的结果进行损失函数计算，与 MLP 网络的结果各自进行反向传播。通过 LSTM 网络类似“监督”的作用，组合网络会从静态样本数据和与时序相关数据的两个角度对时间序列维度间的因果关系进行学习。

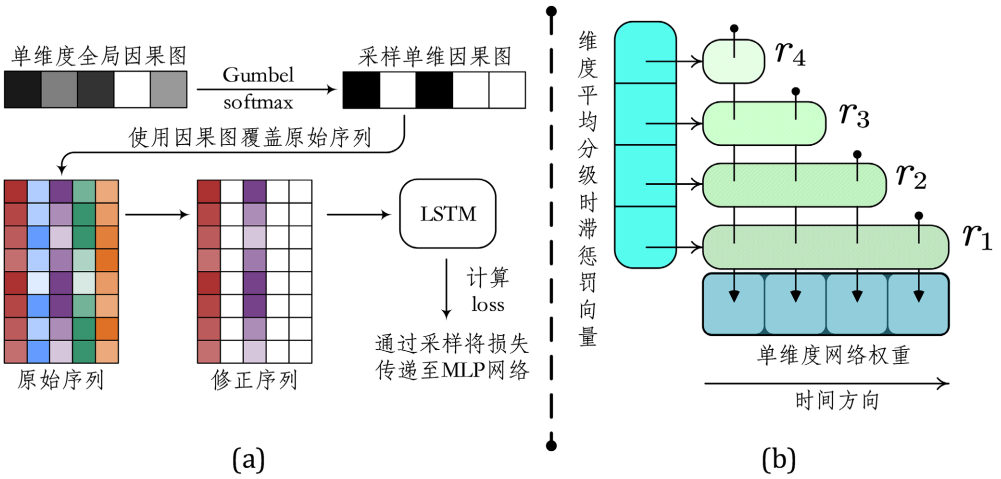


**Figure 4.** (a) Sampling input LSTM process (b) Weighted hierarchical group lasso penalty
**图 4.** (a) 采样输入 LSTM 流程；(b) 加权分级 group lasso 惩罚

### 单维度因果图的伯努利采样

在 MLP 网络中，每一个单维度组件模型提取出的因果图大小为 $G_i\in\mathbb{R}^p$，其中 $i=1,\cdots,p$，多个单维度因果图拼接得到完整因果图 $G$。因果图 $G_i$ 中的项 $c_{ij}$ 指示序列 $j$ 对序列 $i$ 的因果影响程度。使用 Gumbel-

softmax 采样对 $G_i$ 进行优化[25]，即

$$s_{ij}=\frac{e^{\left(\log\left(c_{ij}\right)+g\right)/\tau}}{e^{\left(\log\left(c_{ij}\right)+g\right)/\tau}+e^{\left(\log\left(1-c_{ij}\right)+g\right)/\tau}} \tag{5}$$

其中，$g=-\log\left(-\log\left(u\right)\right)$，$u\sim\text{Uniform}\left(0,1\right)$。$\tau$ 是一个“温度系数”，$\tau$ 值越小，采样就越接近伯努利采样。选择软采样的方式，在进行了 softmax 求期望操作之后，采样接近 1 的维度会保留其原始的序列数据，其对目标维度的影响不变；采样接近 0 的维度序列会被很大程度压缩，其对目标维度的影响在循环网络中也会相应大幅度减少。

使用 Gumbel-softmax 采样方式实现了采样操作可以反向传递，确保 LSTM 网络的损失信息可以回传至 MLP 的输入层权重，对其进行修正。

### 4.3. 网络损失函数与动态稀疏正则化

如图 2 所示，本文提出的网络模型在训练过程中的损失函数由三部分组成，分别是 MLP 网络部分的预测误差、LSTM 网络部分的预测误差和 MLP 网络中输入权重的动态稀疏正则项。其中 MLP 网络的预测误差用来优化因果图提取网络的权重，使其输入权重快速收敛至正确的值，确保可以在数值上提取到小误差的因果图。LSTM 网络的预测误差用于在处理过的输入数据下将循环网络的权重优化至正确的值，使 LSTM 网络可以通过采样的操作来“修正”全局因果图，更快排除无关维度对目标维度的因果影响。添加 MLP 网络输入权重的动态稀疏正则项是为了对输入权重的绝对值进行约束，使更多的权重值可以收敛到精确的 0 值，进而使得真正有因果关系的变量可以获得它们正确的因果关系值。

针对目标变量 $i$，通过最小化式(6)给出的全局损失函数来优化模型

$$\min_{W}\sum_{t=K}^{T}\left(x_{it}-g_i\left(x_{(t-K):(t-1)}\right)\right)^2+\sum_{t=K}^{T}\left(x_{it}-l_i\left(x_{(t-K):(t-1)}\right)\right)+\lambda\rho\left(W^1\right). \tag{6}$$

其中，$g_i\left(\cdot\right)$ 代表目标变量 $i$ 的 MLP 网络部分的近似函数，$l_i\left(\cdot\right)$ 代表目标变量 $i$ 的 LSTM 网络部分的近似函数。$\lambda\rho\left(\cdot\right)$ 代表施加在 MLP 网络输入权重上的带系数正则项。在实际的神经网络中，对 MLP 网络的更高层权重和 LSTM 网络权重施加 2-范数正则化，确保将权重约束在一定范围内。

遵照 4.2.1 节中的表述，MLP 网络的输入权重结构是一个大小为 $H\times p\times K$ 的三维张量，$W^1\in\mathbb{R}^{H\times p\times K}$。通过对该张量在 0 和 1 维度上求 F-范数并进行归一化操作，得到维度平均时滞依赖向量 $\boldsymbol{\lambda}_i=\left(\lambda_{i1},\lambda_{i2},\cdots,\lambda_{ik},\cdots,\lambda_{iK}\right)$，其中 $i=1,\cdots,p$ 且 $k=1,2,\cdots,K$。$\lambda_{ik}$ 代表第 $i$ 个变量对过去第 $k$ 个时间点上的值的平均依赖程度，平均的操作在所有变量的维度上执行。对 $\boldsymbol{\lambda}_i$ 进行分级调整，得到维度平均分级时滞惩罚向量。计算过程由式(7)给出

$$\boldsymbol{\lambda}_i'=\frac{1}{\boldsymbol{\lambda}_i},\quad \boldsymbol{\lambda}_i''=\frac{\boldsymbol{\lambda}_i'}{\lambda_{i1}'},\quad r_{i1}=\lambda_{i1}'',\quad r_{ik}=\lambda_{ik}''-\lambda_{i(k-1)}'',\quad \mathbf{r}_i=\left(r_{i1},\cdots,r_{iK}\right),\quad k=2,\cdots,K \tag{7}$$

其中 $\boldsymbol{\lambda}_i$ 是由网络权重求范数得到的维度平均时滞依赖向量，$\boldsymbol{\lambda}_i'$ 和 $\boldsymbol{\lambda}_i''$ 是中间变量，$\mathbf{r}_i$ 是变量 $i$ 的维度平均分级时滞惩罚向量，用 $\mathbf{r}_i$ 对 MLP 网络输入权重值的正则操作进行校正。

动态分级惩罚策略下的 $\rho\left(W^1\right)$ 规划由式(8)给出

$$\rho\left(W^1\right)=\sum_{j=1}^{p}\sum_{k=1}^{K}r_{ik}\left\|W_j^{1k},\cdots,W_j^{1K}\right\|_F. \tag{8}$$

其中 $W_j^{1k}$ 代表输入变量 $j$ 对应的输入权重在 $k$ 时滞的列，$r_{ik}$ 是 $\mathbf{r}_i$ 的第 $k$ 项，$\left\|\cdot\right\|_F$ 是求 F-范数操作，如图 4(b)所示。

### 4.4. 优化带正则项的损失函数

本文采用迭代软阈值收缩算法[26]对网络的损失函数进行优化，迭代软阈值收缩算法是近端梯度下降算法的一种特殊形式，其损失函数的形式为预测均方误差加上一个带系数的稀疏正则项。该方法驱使参数矩阵的某些行或列陷入 0 阈值范围内，即获得精确的 0 解，这符合本文在 3.2 节中定义的非线性格兰杰因果关系所追求的目标，即发现并排除不存在格兰杰因果关系的两个变量。

在使用收缩算法进行目标优化时，采用线搜索的方式，保证损失函数可以收敛至局部极小值。网络权重由标准正态分布进行随机初始化为 $\boldsymbol{W}^{(0)}$，算法按照式(9)的方式从 $\boldsymbol{W}^{(0)}$ 开始迭代更新网络权重

$$\boldsymbol{W}^{(n+1)} = \text{prox}_{d^{(m)}\lambda\rho}\left(\boldsymbol{W}^{(n)} - d^{(n)}\nabla\mathcal{L}\left(\boldsymbol{W}^{(n)}\right)\right), \tag{9}$$

其中 $\boldsymbol{W}^{(n)}$ 代表第 $n$ 个迭代步的网络权重值。$d^{(n)}$ 代表第 $n$ 个迭代步的步长值。$\mathcal{L}(\boldsymbol{W})$ 是网络的预测误差，由线性和循环两部分构成。$\text{prox}_{\lambda\rho}(\cdot)$ 是关于稀疏惩罚函数 $\rho(\cdot)$ 和正则系数 $\lambda$ 的近似算子。通过线搜索的方式，算法在每一步中驱动预测损失下降以得到“中间值”，再通过修正中间值得到下一步的数值。

在动态正则化策略中，动态稀疏约束只施加于 MLP 网络部分的输入权重上，MLP 网络的其余层和 LSTM 网络的所有层权重的迭代步长都是单一的定值，施加在 MLP 网络输入权重上的动态群组 Lasso 惩罚的近似步长通过对输入权重进行分层加权软阈值运算得到，运算过程由式(10)给出

$$\text{prox}_{d^{(m)}\lambda\rho}\left(W_{:k}^{1}\right) = \left(1 - \frac{r_k \lambda d^{(m)}}{\left\|W_{:k}^{1}\right\|_F}\right)_{+} W_{:k}^{1} \tag{10}$$

其中，$W_{:k}^{1}$ 是输入权重中代表对应前 $k$ 阶滞后值的部分，$r_k$ 是对应每个滞后阶数的平均分级时滞惩罚值。$(\theta)_{+} = \max(0,\theta)$，由于对权重取 F 范数后数值为非负值，所以只需要考虑权重由正值落入阈值范围内的情况。输入权重的近似步长是通过对惩罚函数中每个不同滞后范围的群组迭代应用群组软阈值操作来计算的。

## 5. 实验

本文针对提出的模型在模拟数据集和真实场景启发的数据集上进行验证，将因果图的学习结果与几种主流的基准方法的实验结果进行对比，以验证本文提出方法的有效性。其次，为模型设置不同的正则化参数来检验其对模型训练效果的影响。

### 5.1. 数据集

#### 5.1.1. 向量自回归模型 VAR 数据

VAR 数据[9]是通过设定好的因果图、时滞和预热数据模拟生成的可变长度时间序列数据，我们在 VAR 数据上验证模型在数据因果关系为线性关系时的效果。模拟序列维度为 $p=10$ 的线性 VAR(2)和 VAR(3)数据。需要指出的是，VAR($n$)数据集指明了在数据生成时当前时刻数据与过去的 $n$ 个时刻值相关。如果变量 $i$ 依赖于变量 $j$，则设置因果图矩阵中对应项 $A_{ij}^{k}=0.1$，其中 $k=1,2,3$。因果图矩阵中的其他项均为零，标识对应变量之间无依赖关系。为了验证 DRGC 对于不同时间长度序列的有效性以及 LSTM 网络对时间信息捕捉的能力，分别生成长度为 200、500 和 1000 个时间点的序列数据。

#### 5.1.2. Lorenz-96 数据

Lorenz-96 数据集是一个在混沌理论中广泛使用的合成数据集，用于研究非线性动力系统的复杂行为[27]。该数据集通过一个简化的三维流体动力学模型生成，尽管其维度较低，但却能够展现出高度复杂的

混沌动态。$p$ 维连续的 lorenz-96 模型的变化方式由式(11)给出

$$\frac{\mathrm{d}x_{ti}}{\mathrm{d}t}=\left(x_{t(i+1)}-x_{t(i-2)}\right)x_{t(i-1)}-x_{ti}+F \tag{11}$$

其中 $i=1,\cdots,p$，为了使得式(11)对所有的 $i$ 值都有意义，有式(12)给出的定义。

$$x_{t(-1)}=x_{t(p-1)},\ x_{t0}=x_{tp},\ x_{t(p+1)}=x_{t1} \tag{12}$$

通过式(12)的定义，变量形成一个循环链。$F$ 是一个“强迫常数”，决定了序列非线性的程度和“混乱”的程度。本文模拟 $p=10$ 的 Lorenz-96 数据并采用两种不同的强迫常数，采样率设置为 $\Delta t=0.05$，得到具有稀疏格兰杰因果关系的多变量非线性时间序列。

#### 5.1.3. DREAM-3 数据

我们在真实场景启发的时间基因表达数据集 DREAM-3 [28]上验证提出的模型效果。该数据集在格兰杰因果关系检验上是一个较困难的非线性数据集。数据使用连续的基因表达和调控动力学来模拟，背后有多个隐藏因素是未被观察到的。数据共包含五个有着不同真实因果图的模拟数据集。每个数据集都包含 10 节点数据的 4 段序列，每段序列均有 21 个采样时间点，将不同序列前后拼接构成时间长度为 84 的总序列。

### 5.2. 基准方法

本文选择了 4 种主流的基准方法进行比较：1) 神经格兰杰因果发现方法 NGC [4]，它利用多层感知机和循环神经网络两种组件模型结合群组惩罚来推断格兰杰因果关系。在本文实验中针对 VAR 数据集使用 MLP 模型，针对 Lorenz-96 数据使用 RNN 模型。2) PCMCI [29]，一种使用条件独立测试来检测非线性格兰杰因果关系的方法。3) economy-SRU [5]，一种使用基于统计回归单位(Statistical Regression Units, SRU)的分量时间序列预测模型进行非线性建模的方法，通过设计少量的可训练参数，提高了模型对预测数据的抗过拟合性能。4) CUTS [22]，一种通过交替进行因果发现和潜在数据预测两个阶段来学习数据中的因果关系的方法，这种方法可以对不规则数据(存在缺失值的时间序列)同时进行因果发现和数据填充。它同样适用于不存在缺失值的时间序列数据集上，并且反映出较好的学习结果。

### 5.3. 模拟数据实验结果

在定量评估方面，为了验证因果图复原数据底层结构的准确性，实验中以 ROC 曲线下面积指标(Area Under the Receiver Operating Characteristic Curve, AUROC)作为标准。具体来说，对本文提出的方法和 NGC 方法，通过将一定范围内变化的 $\lambda$ 值作为动态正则项的系数进行模型训练，得到模型对数据集上无阈值因果图学习的 ROC 曲线，并计算 AUROC 值；对 PCMCI 和 eSRU 方法则是使用不同的阈值来获取 AUROC 值。

#### 5.3.1. 因果图学习的 AUROC 比较

本部分实验生成 VAR(1)、VAR(2)、VAR(3)数据和使用两种不同“强迫常数”的 Lorenz-96 数据，每个数据集都有 200、500 和 1000 三个不同时间长度的版本，以观察在数据集长度变化的情况下各个方法的学习效果，如表 1 所示。表中的 DRGC-s 方法代表在本文提出的模型基础上除去循环网络模块的消融模型。表 1 中所呈现的数值(如 99.86 ± 0.12)均为 AUROC 的数值表示，其完整数值为对应的 0.9986 ± 0.0012 (若换算为小数形式)或者 98.66% ± 0.12% (若换算为百分比形式)。

**Table 1.** AUROC for causal graph using different methods on 4 time series datasets
**表 1.** 不同方法在 4 个时间序列数据上因果图学习的 AUROC 对比

| Model | VAR(2) | | | VAR(3) | | |
|---|---|---|---|---|---|---|
| T | 200 | 500 | 1000 | 200 | 500 | 1000 |
| DRGC | **99.86 ± 0.12** | **100.00 ± 0.00** | **100.00 ± 0.00** | **98.86 ± 1.42** | **100.00 ± 0.00** | **100.00 ± 0.00** |
| DRGC-s | 99.17 ± 0.98 | 100.00 ± 0.00 | 100.00 ± 0.00 | 97.56 ± 2.05 | 100.00 ± 0.00 | 100.00 ± 0.00 |
| NGC | 96.01 ± 2.14 | 91.98 ± 3.87 | 98.52 ± 1.31 | 90.91 ± 4.15 | 97.53 ± 2.82 | 98.74 ± 0.98 |
| PCMCI | 71.40 ± 4.93 | 72.30 ± 5.73 | 72.19 ± 3.25 | 65.89 ± 5.23 | 71.30 ± 4.85 | 71.92 ± 3.61 |
| eSRU | 87.74 ± 6.71 | 89.97 ± 5.72 | 91.06 ± 4.86 | 81.25 ± 7.05 | 86.55 ± 5.48 | 90.38 ± 2.29 |
| CUTS | 98.92 ± 0.84 | 99.05 ± 1.03 | 100.00 ± 0.00 | 98.05 ± 1.64 | 99.91 ± 0.03 | 99.99 ± 0.02 |
| Model | Lorenz-96 (F= 10) | | | Lorenz-96 (F = 20) | | |
| T | 200 | 500 | 1000 | 200 | 500 | 1000 |
| DRGC | 94.91 ± 0.96 | **100.00 ± 0.00** | **100.00 ± 0.00** | **92.00 ± 0.97** | 95.49 ± 2.30 | **98.80 ± 0.45** |
| DRGC-s | 93.63 ± 0.55 | 98.64 ± 0.86 | 100.00 ± 0.00 | 90.51 ± 1.48 | **95.66 ± 1.62** | 98.01 ± 1.57 |
| NGC | 93.88 ± 1.64 | 98.56 ± 0.51 | 99.15 ± 0.33 | 84.35 ± 3.31 | 92.22 ± 3.72 | 92.82 ± 2.60 |
| PCMCI | 85.95 ± 4.55 | 91.65 ± 3.54 | 95.72 ± 2.23 | 80.87 ± 5.21 | 86.56 ± 3.18 | 86.39 ± 2.71 |
| eSRU | 87.42 ± 3.98 | 96.45 ± 3.06 | 97.51 ± 1.16 | 90.30 ± 3.63 | 97.57 ± 1.43 | 98.43 ± 1.74 |
| CUTS | **95.12 ± 1.13** | 99.64 ± 0.51 | 100.00 ± 0.00 | 89.20 ± 0.73 | 93.36 ± 0.82 | 97.15 ± 2.94 |

实验结果表明，DRGC 在不同长度的 VAR 数据和 Lorenz-96 数据上基本取得了最优的结果。当时间序列的长度越长时，提供给模型的训练样本越多，所以基本所有方法在长度为 500 和 1000 数据上的结果都优于长度为 200 的数据。此外，由于 VAR 数据结构较为简单，DRGC 可以完全准确地识别出系统中每个变量的真实原因变量，并在混乱度较低的 Lorenz 数据上保持优秀的识别率。相比于 NGC 方法，DRGC 方法在长序列上大致有 1%至 5%的指标提升，这是动态正则化的作用。滞后值的正确选择使得模型能够准确量化目标变量对原因变量的依赖程度，进而更为精确地排除无关变量的微小影响，事实上，NGC 方法有时选择出的错误因果关系相比于正确的那些在任何滞后值上都只有很小的权重，而传统的正则方法无法很好地排除它们的影响。

### 5.3.2. 不同依赖系数的 VAR(3)数据上的 AUROC 比较

VAR 数据的依赖系数 $p$ 代表了在这个模拟系统中每个变量受几个变量的驱动，$p = 0.2$ 代表驱动目标变量变化的“原因”变量个数占系统中总变量个数的 20%，同时每个目标变量的“原因”变量一定包含其本身。$p$ 值越大，系统中相互依赖的变量数量就越多，对应的因果图边的数量也越多，系统变得复杂，模型发现正确的因果图则更加困难。本部分实验生成依赖系数分别为 0.2、0.3 和 0.4 的三种 VAR(3)数据，且同样有三种不同的时间长度，以验证模型在面对不同复杂程度的数据集时发现因果图的能力。学习结果如表 2 所示，表中数据同样是百分比数值。

**Table 2.** AUROC for causal graph on VAR(3) with three dependency coefficients $p$
**表 2.** 不同方法在 3 种依赖系数 $p$ 下的 VAR(3)数据上因果图学习的 AUROC 对比

| Model | VAR(3) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $p$ = 0.2 | | | $p$ = 0.3 | | | $p$ = 0.4 | | |
| T | 200 | 500 | 1000 | 200 | 500 | 1000 | 200 | 500 | 1000 |
| DRGC | **98.86 ± 1.42** | **100.0 ± 0.00** | **100.0 ± 0.00** | **86.57 ± 6.07** | 90.63 ± 5.08 | 94.14 ± 5.34 | **78.67 ± 7.13** | 85.82 ± 2.27 | 88.21 ± 6.77 |
| DRGC-s | 97.56 ± 2.05 | 100.0 ± 0.00 | 100.0 ± 0.00 | 85.64 ± 4.09 | **99.43 ± 0.75** | **100.0 ± 0.00** | 76.06 ± 5.96 | **91.99 ± 1.38** | **95.72 ± 0.51** |
| NGC | 90.91 ± 4.15 | 97.53 ± 2.82 | 98.74 ± 0.98 | 81.18 ± 3.16 | 90.38 ± 3.91 | 93.84 ± 3.27 | 75.85 ± 5.02 | 82.84 ± 4.55 | 87.82 ± 3.67 |
| PCMCI | 65.89 ± 5.23 | 71.30 ± 4.85 | 71.92 ± 3.61 | 57.52 ± 4.83 | 54.95 ± 6.50 | 54.91 ± 5.81 | 56.49 ± 7.42 | 51.95 ± 7.75 | 53.49 ± 7.78 |
| eSRU | 81.23 ± 7.16 | 86.22 ± 5.49 | 90.10 ± 2.66 | 70.75 ± 7.33 | 75.68 ± 6.73 | 82.26 ± 4.11 | 63.95 ± 7.48 | 68.60 ± 5.13 | 73.99 ± 8.05 |
| CUTS | 98.97 ± 1.64 | 99.91 ± 0.03 | 99.99 ± 0.02 | 82.58 ± 4.79 | 95.33 ± 0.99 | 94.45 ± 2.28 | 75.91 ± 6.73 | 83.43 ± 1.65 | 92.59 ± 3.31 |

实验结果表明，随着依赖系数增大，时间序列系统各变量间的因果关系更为复杂，各个方法对因果关系学习的准确程度有所下降。当系统中每个变量仅有两个依赖变量时，DRGC 方法可以准确预测出正确的因果图。对于较复杂的系统因果图，DRGC 方法仍能领先其他方法表达出较好的效果。序列长度为 500 和 1000 时，得益于足够多的训练样本，消融模型学习到了最好的结果；而带有 LSTM 网络的完整模型则能够在时间序列长度较短时更好地捕捉时间维度上的依赖关系，表现出更好的性能。

### 5.3.3. 在 VAR(3)数据上与 NGC 的时滞选择比较

VAR(3)数据在模拟生成时，每一时刻某变量的数据值仅受其原因变量过去最多三个时刻的值的影响，即 $x_t$ 与 $x_{t-4}$ 无因果关系。DRGC 方法中的线性网络动态正则策略的最大优势在于可以定量还原不同滞后值之间的相对因果关系大小，同时不影响对无关变量的排除。在生成 VAR 数据时，本文设置当前值对过去 3 个滞后值的依赖程度相同，体现在因果图上则是具有相同的权重值。为了检验模型排除高阶滞后值的能力，在线性网络一维卷积时使用的卷积核大小为 5。选取 10 维中的第 2 个和第 5 个变量为例，观察其时滞因果图的学习结果，如图 5 所示。第 2 变量的原因变量是第 2 和第 5 变量，第 5 变量的原因变量是第 5 和第 7 变量，两种方法学习到的单变量对单变量的时滞差异平均值的对比如表 3 所示。

**Table 3.** Mean value of the difference in the delay of a single variable on VAR(3)
**表 3.** 在 VAR(3)上单一变量时滞差异平均值对比

| Model | 变量 2-变量 2 | 变量 2-变量 5 | 变量 5-变量 5 | 变量 5-变量 7 |
|---|---|---|---|---|
| cMLP | 0.47255 | 0.30196 | 0.43922 | 0.46863 |
| DRGC | 0.09412 | 0.06564 | 0.11765 | 0.19216 |

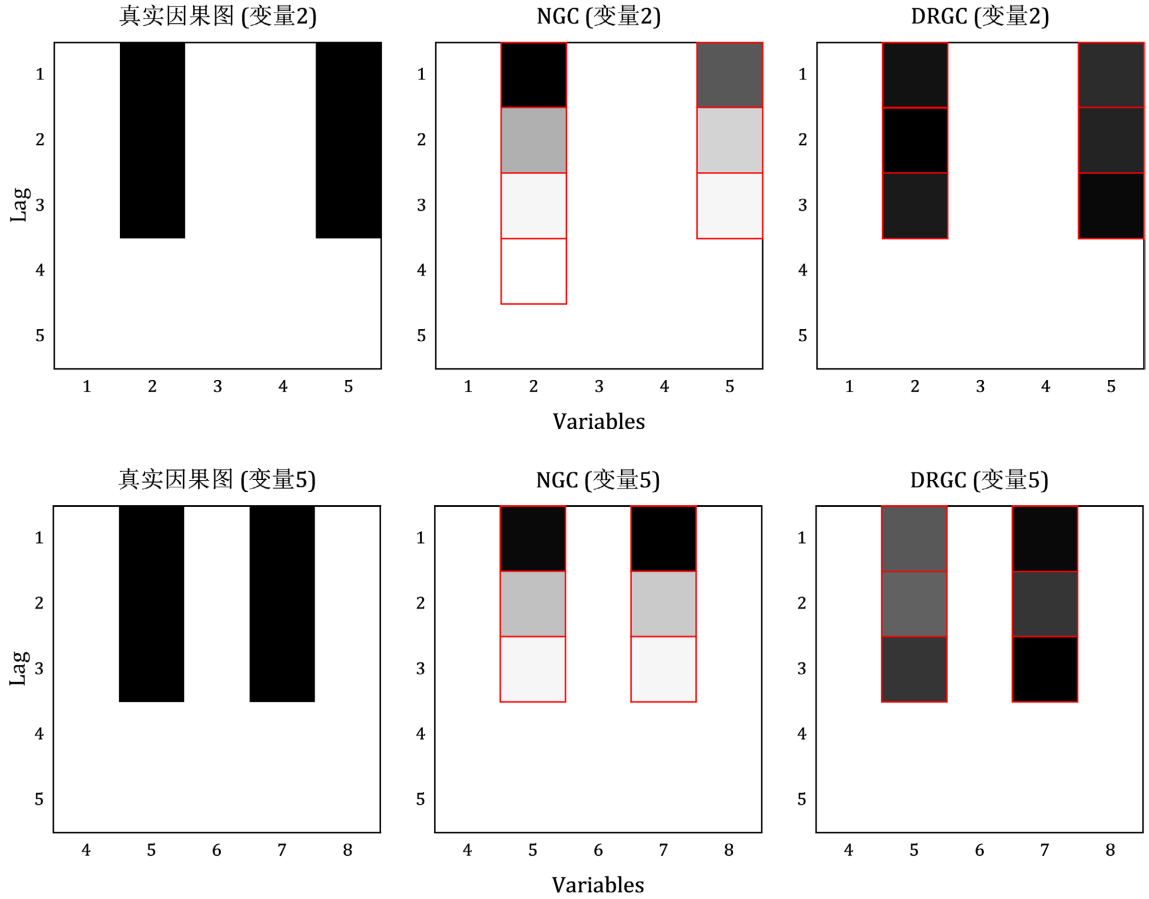


**Figure 5.** Lagged causal graph on VAR(3) of DRGC and cMLP
**图 5.** VAR(3)上的滞后因果图学习结果

由图 5 可得，cMLP 的学习结果由于受到硬性的分级递增惩罚策略的限制，在选取的两个变量上的因果关系大小均随着滞后值的增大而有非常明显的减弱。同时当预设的最大滞后阶数增加时，惩罚的变化较为缓慢，不能较好地排除无关滞后值的影响。DRGC 方法学习到了目标变量在不同滞后值上相对差异较小的因果关系大小，这符合 VAR(3)数据在模拟生成时“过去 3 个时刻值具有相同大小的因果关系”的正确机制。

从具体数值的差异方面来看，DRGC 方法对各个变量时滞权重之间的差异相比 cMLP 缩减了 20%至 40%。表 3 中所示的差异平均值计算方法是对两个单变量之间的三阶时滞值作差，然后对两个差值求平均，所有的时滞值在同一个目标变量下进行了归一化处理。DRGC 方法将时滞因果强度之间的差异缩小到了 10%，而 cMLP 方法的平均差异处于 40%至 50%之间，相当于最大时滞因果强度的一半。由于 cMLP 采用按时滞递增的惩罚方法，相邻时滞之间的惩罚差异相同，会导致学习到的时滞在一定程度上呈现出均匀下降的趋势。DRGC 方法实现了最后一阶时滞与无因果时滞之间惩罚值的跳变，可以更加合理地拟合时滞的下降趋势。

#### 5.3.4. 不同正则系数下 VAR(3)数据的时滞选择

DRGC 方法在线性层的输入权重上施加了动态的分级稀疏惩罚。对于对应各组滞后的输入权重，除了需要加入其维度平均分级时滞惩罚的值外，还需要为所有的惩罚值设置一个统一的系数 $\lambda$，以实现神经网络对各个变量和各个滞后值对应权重的整体调控。当 $\lambda$ 的值变化时，模型对数据的因果图提取结果

也会变化。

本部分实验使用与 5.3.3 节中相同设置的 VAR 数据集，采用多种不同的动态正则系数 $\lambda$ 进行时滞选择实验并与真实的滞后因果图进行对比，我们选择了其中一个对比较为明显的变量进行展示。

图 6 展示了一系列动态正则系数下模型对目标变量包含滞后因果图的学习结果和其真实的因果图，我们训练一个具有分级 Lasso 惩罚和最大滞后阶数为 5 的 cMLP 模型作为对比。

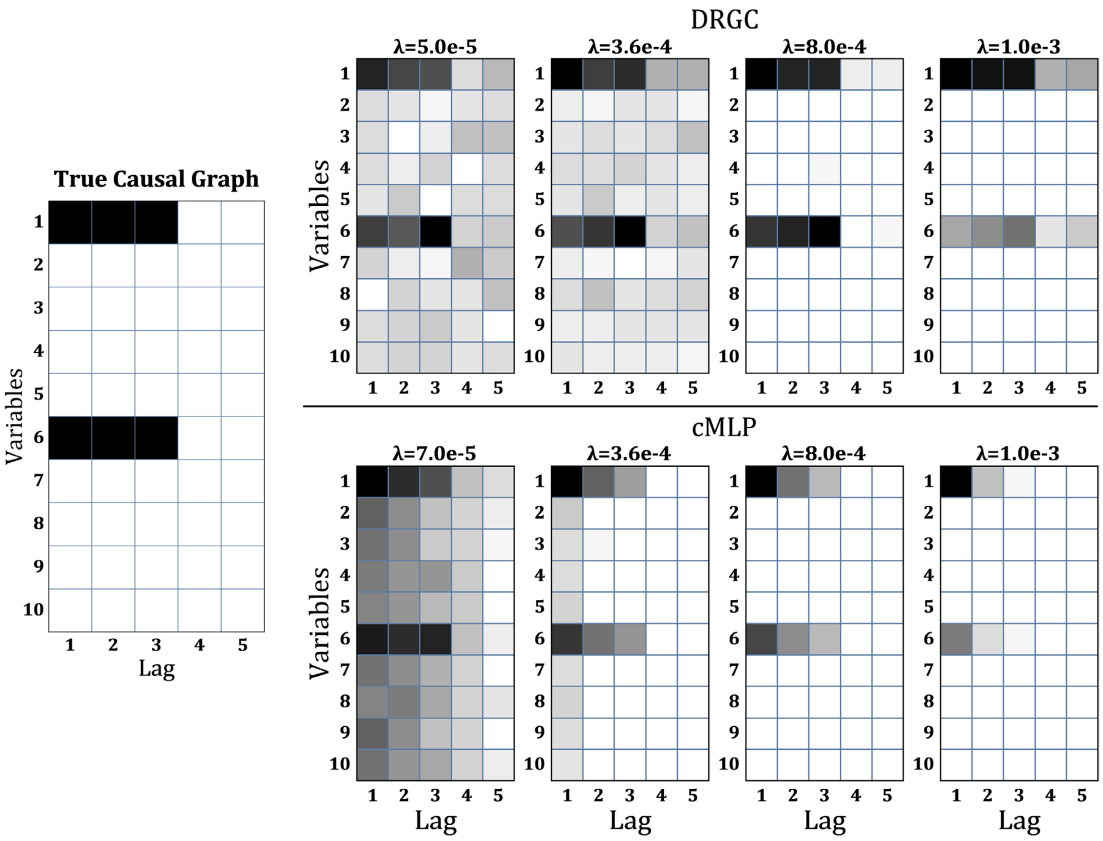


**Figure 6.** Lagged causal graph of DRGC and cMLP with different dynamic regularization coefficients
**图 6.** 不同动态正则系数下 DRGC 和 cMLP 的时滞因果图学习结果

由图 6 可得，当 $\lambda$ 取值适当时，NGC 和 DRGC 方法都可以很好地发现目标变量正确的因果图，且 DRGC 相较于 NGC 可以更合理地分配不同滞后值的因果大小。对于较小的 $\lambda$ 值，NGC 方法会过多估计原因变量的滞后阶数；而 DRGC 方法则可以更多地保持住不同滞后值之间的相对平稳，在无关的变量上的估计值相对很小，可以通过设置相应的阈值来排除影响。对于较大的 $\lambda$ 值，NGC 方法会过多削弱更靠后的滞后值的大小，而 DRGC 方法仍可以较好地保持各滞后值的相对大小关系，同时更多地排除无关变量的影响。当 $\lambda$ 值设置得过高时，由于输入权重惩罚值与预测误差之间的平衡，模型将学习不到目标变量的因果图，权重值被惩罚至 0。实验中可以通过交叉验证来选择合适的 $\lambda$ 值。

## 5.4. DREAM-3 实验

DREAM-3 数据集是用于严格比较格兰杰因果关系检测方法的较困难的非线性数据集，包含三个 Yeast 数据和两个 E. Coli 数据。我们分别在这 5 个数据集上使用 DRGC 方法进行学习，同时选用 NGC 方法的线性模型和循环模型分别作为对比。由于 DREAM-3 数据的序列长度较短，且数据中只有真实的全

局因果图，我们将模型中的最大滞后阶数设置为 2，隐藏层单元个数设置为 10，以减轻模型的体量并加快训练过程，并对 NGC 方法采用相同的设置。在指标方面，仍使用 AUROC 来评估因果图的学习结果，在 5 个时间序列数据集上的评估结果如表 4 所示，对应的 ROC 曲线如图 7 所示。

**Table 4.** ROC for DRGC, cMLP, and cLSTM models on DREAM-3 datasets
**表 4.** DRGC，cMLP 和 cLSTM 在 5 个 Dream 数据集上的 AUROC 对比

| Model | Yeast1 | Yeast2 | Yeast3 | Ecoli1 | Ecoli2 |
|---|---|---|---|---|---|
| cMLP | 0.5722 | 0.5629 | 0.5553 | 0.5741 | 0.5855 |
| cLSTM | 0.5933 | 0.58 | 0.5525 | 0.5132 | 0.5647 |
| DRGC | **0.6568** | **0.6485** | **0.6067** | **0.6236** | **0.6070** |

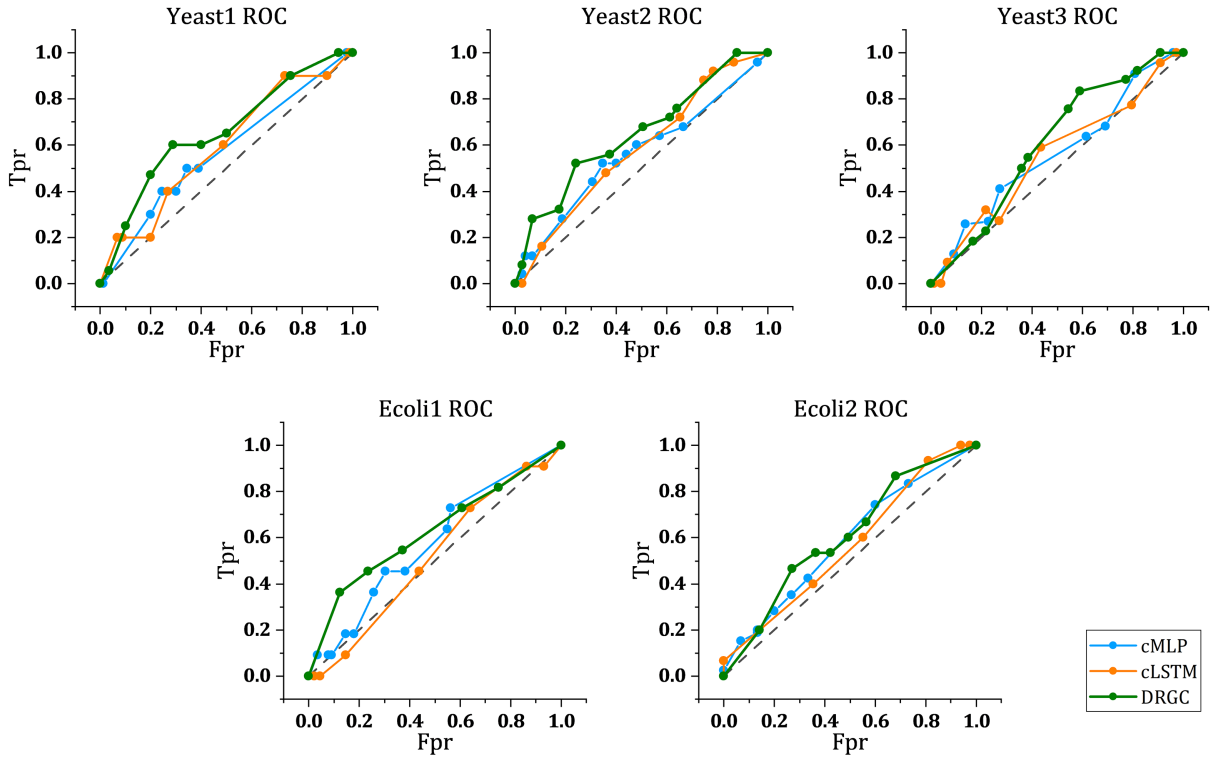


**Figure 7.** ROC for DRGC, cMLP, and cLSTM models on DREAM-3 datasets
**图 7.** DRGC，cMLP 和 cLSTM 在 Dream 数据集上的 ROC 曲线

由表 4 可得，DRGC 方法在所有五个数据集上的结果优于 cMLP 和 cLSTM 两个模型，由于 10 节点的 Dream 数据集只包含 4 个长度为 21 的时间序列，用于训练的样本数量较少，模型学习到正确因果图的难度较高，各个模型的 AUROC 值均较低，但 DRGC 方法达到了 60%的指标。DRGC 方法融合了线性网络与循环网络的优点，可以使权重值向对应真实原因变量的单元集中。

## 6. 结论

本文提出了一种采用动态分级稀疏惩罚策略的线性与循环网络组合多维时间序列非线性格兰杰因果挖掘方法。为提高模型的可解释性，我们为系统中的每个变量单独建立线性网络和采样输入循环网络，

依靠循环网络提取时间方向依赖关系的能力，对线性网络挖掘的整体因果关系进行监督和修正。为了提高模型在不同时滞上选择因果关系的准确性，通过从线性网络输入权重中提取变量对不同时滞的平均依赖程度，对模型施加动态变化的稀疏惩罚。在模拟数据集和真实基因调控子网络生成的数据集上进行的实验结果表明，DRGC 方法在不同长度和复杂程度的数据集上都表现出领先于其他现有方法的性能。

未来的工作方向有两点，1) 探究模型如何在脱离数据集的真实全局因果关系和时间因果关系信息的情况下，配合因果图的阈值设置来进行更合理的参数调整与选取；2) 在更多的真实世界数据集上进行实验，探寻合适的底层系统结构来协助进行因果关系验证，以提高方法的可推广性。

## 基金项目

国家自然科学基金(62262016)；

中国澳门基金会 2024 年学术资助计划“建设横琴数据跨境传输安全管理试点(数据飞地)的多视角可行政策研究”(G01156-2309-262)；

中央高校基本科研业务费专项资金(2023JBZY035)资助。

## 参考文献

[1] Vicente, R., Wibral, M., Lindner, M. and Pipa, G. (2010) Transfer Entropy—A Model-Free Measure of Effective Connectivity for the Neurosciences. *Journal of Computational Neuroscience*, **30**, 45-67. https://doi.org/10.1007/s10827-010-0262-3

[2] Runge, J. (2018) Causal Network Reconstruction from Time Series: From Theoretical Assumptions to Practical Estimation. *Chaos*: *An Interdisciplinary Journal of Nonlinear Science*, **28**, Article ID: 075310. https://doi.org/10.1063/1.5025050

[3] Gerhardus, A. and Runge, J. (2020) High-Recall Causal Discovery for Autocorrelated Time Series with Latent Confounders. *Advances in Neural Information Processing Systems* 33: *Annual Conference on Neural Information Processing Systems* 2020, *NeurIPS* 2020, 6-12 December 2020, 12615-12625.

[4] Tank, A., Covert, I., Foti, N., Shojaie, A. and Fox, E.B. (2021) Neural Granger Causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 4267-4279. https://doi.org/10.1109/tpami.2021.3065601

[5] Khanna, S. and Tan, V.Y.F. (2019) Economy Statistical Recurrent Units for Inferring Nonlinear Granger Causality.

[6] Pamfil, R., Sriwattanaworachai, N., Desai, S., *et al*. (2020) Dynotears: Structure Learning from Time-Series Data. *International Conference on Artificial Intelligence and Statistics*, 26-28 August 2020, 1595-1605.

[7] Granger, C.W.J. (1969) Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, **37**, 424-438. https://doi.org/10.2307/1912791

[8] Marinazzo, D., Pellicoro, M. and Stramaglia, S. (2008) Kernel-Granger Causality and the Analysis of Dynamical Networks. *Physical Review E*, **77**, Article ID: 056215. https://doi.org/10.1103/physreve.77.056215

[9] Lütkepohl, H. (2005) New Introduction to Multiple Time Series Analysis. Springers Science & Business Media.

[10] Lusch, B., Maia, P.D. and Kutz, J.N. (2016) Inferring Connectivity in Networked Dynamical Systems: Challenges Using Granger Causality. *Physical Review E*, **94**, Article ID: 032220. https://doi.org/10.1103/physreve.94.032220

[11] Amblard, P. and Michel, O.J.J. (2010) On Directed Information Theory and Granger Causality Graphs. *Journal of Computational Neuroscience*, **30**, 7-16. https://doi.org/10.1007/s10827-010-0231-x

[12] Yu, R., Zheng, S., Anandkumar, A., *et al*. (2018) Long-Term Forecasting Using Tensor-Train RNNs.

[13] Li, Y., Yu, R., Shahabi, C., *et al*. (2017) Graph Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting.

[14] Lozano, A.C., Abe, N., Liu, Y. and Rosset, S. (2009) Grouped Graphical Granger Modeling Methods for Temporal Causal Modeling. *Proceedings of the* 15*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, 28 June-1 July 2009, 577-586. https://doi.org/10.1145/1557019.1557085

[15] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[16] Lozano, A.C., Abe, N., Liu, Y. and Rosset, S. (2009) Grouped Graphical Granger Modeling for Gene Expression Regulatory Networks Discovery. *Bioinformatics*, **25**, i110-i118. https://doi.org/10.1093/bioinformatics/btp199

[17] Runge, J., Heitzig, J., Petoukhov, V. and Kurths, J. (2012) Escaping the Curse of Dimensionality in Estimating Multivariate Transfer Entropy. *Physical Review Letters*, **108**, Article ID: 258701. https://doi.org/10.1103/physrevlett.108.258701

[18] Wu, T., Breuel, T., Skuhersky, M., *et al*. (2020) Discovering Nonlinear Relations with Minimum Predictive Information Regularization.

[19] Xu, C., Huang, H. and Yoo, S. (2019) Scalable Causal Graph Learning through a Deep Neural Network. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, 3-7 November 2019, 1853-1862. https://doi.org/10.1145/3357384.3357864

[20] Singh, R., Wu, A.P. and Berger, B. (2022) Granger Causal Inference on DAGs Identifies Genomic Loci Regulating Transcription.

[21] Absar, S., Wu, Y. and Zhang, L. (2023) Neural Time-Invariant Causal Discovery from Time Series Data. 2023 *International Joint Conference on Neural Networks* (*IJCNN*), Gold Coast, 18-23 June 2023, 1-8. https://doi.org/10.1109/ijcnn54540.2023.10192004

[22] Cheng, Y., Yang, R., Xiao, T., *et al*. (2023) Cuts: Neural Causal Discovery from Irregular Time-Series Data.

[23] Cheng, Y., Li, L., Xiao, T., Li, Z., Suo, J., He, K., *et al*. (2024) CUTS+: High-Dimensional Causal Discovery from Irregular Time-Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 11525-11533. https://doi.org/10.1609/aaai.v38i10.29034

[24] Sultan, M.S., Horvath, S. and Ombao, H. (2022) Granger Causality Using Neural Networks.

[25] Jang, E., Gu, S. and Poole, B. (2016) Categorical Reparameterization with Gumbel-Softmax.

[26] Zhang, J. and Ghanem, B. (2018) ISTA-Net: Interpretable Optimization-Inspired Deep Network for Image Compressive Sensing. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1828-1837. https://doi.org/10.1109/cvpr.2018.00196

[27] Lorenz, E.N. (1996) Predictability: A Problem Partly Solved. *Proceedings on Seminar on Predictability*, **1**, 1-18.

[28] Prill, R.J., Marbach, D., Saez-Rodriguez, J., Sorger, P.K., Alexopoulos, L.G., Xue, X., *et al*. (2010) Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. *PLOS ONE*, **5**, e9202. https://doi.org/10.1371/journal.pone.0009202

[29] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. and Sejdinovic, D. (2019) Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets. *Science Advances*, **5**, eaau4996. https://doi.org/10.1126/sciadv.aau4996