Expert Systems With Applications 207 (2022) 118036

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Deep recurrent modelling of Granger causality with latent confounding

Zexuan Yin *, Paolo Barucca

Department of Computer Science, University College London, WC1E 7JE, United Kingdom

ARTICLE INFO

Keywords:
Latent confounders
Recurrent neural networks
Time series prediction

ABSTRACT

Inferring causal relationships in observational time series data is an important task when interventions cannot be performed. Granger causality is a popular framework to infer potential causal mechanisms between different time series. The original definition of Granger causality is restricted to linear processes and leads to spurious conclusions in the presence of a latent confounder. To this end, we propose a deep learning model to detect nonlinear Granger causality and directly account for latent confounders. Our approach consists of two components: 1. feed-forward neural networks to infer representations of the confounder from available proxy variables; 2. recurrent neural networks to construct forecasting models for the target time series with and without additional information. Conditioned on the proxy, if the target time series can be better predicted without extra information, our model concludes that the confounder alone Granger causes the target, and vice versa. To assess the proposed approach, we tested the model on both synthetic and real world time series with known causal relationships; results showed the superiority of our model relative to existing benchmarks.

1. Introduction

Identifying causal relationships from time series data is important as it helps to facilitate informed decision making. When controlled experiments are feasible, interventions are often performed to break the symmetry of association and provide the direction of causal mechanisms (Eichler, 2012), e.g. predicting patients' response to certain treatments over time (Bica et al., 2020). In reality, causal inference through interventions is not always feasible as it could be unethical, costly, or simply impossible to carry out, as in the case of financial time series (Hiemstra & Jones, 1994) and climate variables (Stips et al., 2016); in those scenarios we resort to causal inference from observational data.

Granger causality (Granger, 1969) is a commonly used framework to infer potential causal relationships. The notion of Granger causality relies on two fundamental principles: 1. the cause precedes the effect in time and 2. the cause contains unique information about the effect not available elsewhere (Eichler, 2012). A time series Granger causes another if its past helps to predict the future values of the target time series (Granger, 1969). Traditionally, model-based Granger causality has been tested mostly on linear dynamics in the form of a vector autoregressive model (VAR) (Yuan & Shou, 2020), where one regresses the lagged values of potential causes against the future value of the target series and assess whether the coefficients are statistically different from zero.

Since real world temporal dynamics are rarely linear, several adaptations to model nonlinear causal relationships have been made using for example polynomial autoregression models (Bezruchko et al., 2008) and kernel-based methods (Marinazzo et al., 2011). Model-free approaches such as transfer entropy (Vicente et al., 2011) are able to detect nonlinear dependencies between time series, however they suffer from high variance and require large amounts of data for reliable estimation (Tank et al., 2021). In this work, we follow a recent trend that uses neural networks to infer complex nonlinear causal dependencies in time series data (Bussmann et al., 2020; De Brouwer et al., 2020; Khanna & Tan, 2020; Marcinkevičs & Vogt, 2021; Moraffah et al., 2021; Nauta et al., 2019; Rahimi et al., 2020; Tank et al., 2021; Trifunov et al., 2019).

An important consideration for causal inference from observational time series is confounding bias. A confounder variable affects both cause and effect and therefore must be accounted for to avoid spurious conclusions. Granger causality relies on the causal sufficiency (no latent confounding) assumption (Spirtes & Zhang, 2016) and is known to be biased in the presence of confounding (Peters et al., 2017). Consider the case where the confounder Z affects the cause variable X with lag 2 and the effect variable Y with lag 4, assuming causal sufficiency would lead to the biased conclusion that X Granger causes Y. When all confounders are observed, the multivariate conditional Granger causality tests can be applied (Chen et al., 2006), which relies on the fact that all variables that could have had a possible influence have been considered in the analysis (Marinazzo et al., 2011). In reality, it is rarely possible to measure all the confounders, nevertheless, we may have

* Corresponding author.
E-mail addresses: zexuan.yin.20@ucl.ac.uk (Z. Yin), p.barucca@ucl.ac.uk (P. Barucca).

https://doi.org/10.1016/j.eswa.2022.118036
Received 6 August 2021; Received in revised form 24 May 2022; Accepted 30 June 2022
Available online 4 July 2022

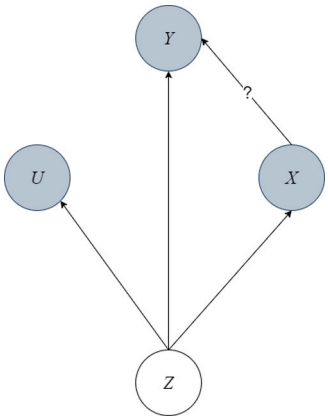

Fig. 1. Causal graph showing the relationship between effect variable Y, cause variable X, latent confounder Z and proxy variable U.

access to noisy measurements of proxies for the confounders (Louizos et al., 2017; Pearl, 2010). Since the majority of existing works on neural network-based approaches to Granger causality assume causal sufficiency, how best to account for latent confounders is still an open question. In this work, neural networks are used to infer representations of the latent confounder from the available proxies, which can be used in the subsequent Granger causality tests.

Consider the causal system in Fig. 1 involving a cause variable $X \in \mathbb{R}^{1\times T}$, an effect variable $Y \in \mathbb{R}^{1\times T}$, a latent confounder $Z \in \mathbb{R}^{1\times T}$ and proxies of the confounder $U \in \mathbb{R}^{n\times T}$, where T is the length of the time series and n is the number of proxies available. Our aim is to infer the Granger causal relationship between the confounded pair X and Y.

Our proposed architecture is a maximum likelihood based latent variable model to learn useful information about the confounder Z from available proxies U, and to model the relationship between Z, X and Y. In practice, the proxy variables are often chosen using expert judgement. Consider a situation where the latent confounder is the socio-economic status of a patient, one could use the zip code or job type of the patient as proxy variables (Louizos et al., 2017). More formally, we follow the assumption in Louizos et al. (2017) that the joint distribution $P(Y, X, Z, U)$ can be approximately recovered from the observations (Y, X, U), which could turn out to be impossible if the confounder has no relation to the observed variables. In this work we focus on the case where: 1. proxy variables are available in abundance to allow recovery of the joint distribution, 2. expert judgement is in place to select appropriate proxies, and 3. (Y, X, U) are potentially complex but learnable functions of Z which we approximate with neural networks. This scenario is termed a "surrogate-rich setting" in Louizos et al. (2017). The main contributions of this work are as follows:

1. we propose a deep learning-based test for nonlinear Granger causality with latent confounding. Multilayer perceptrons are used to infer representations of the confounder Z from available proxy variables U. Recurrent neural networks are used to parameterise two forecasting models (predictive distributions) for Y, with and without X. A two-sample t-test is performed to establish whether the inclusion of X results in a statistically significant reduction in the prediction error of Y, and hence a Granger causal relationship $X \rightarrow Y$.
2. we propose the novel use of a dual-decoder setup corresponding to the two predictive distributions mentioned above. This avoids the need to train two separate neural networks for comparison of predictive accuracy
3. we demonstrate the effectiveness of the proposed approach on datasets with known data generating processes, and sensitivity analyses were conducted to show the robustness of the model

2. Related work

The original definition of Granger causality (Granger, 1969) involves linear dynamics studied using a VAR model. For a collection of k time series $\boldsymbol{X} \in \mathbb{R}^{k\times T}$ and $\boldsymbol{X}_t \in \mathbb{R}^k$ a VAR model is defined:

$$\boldsymbol{X}_t = \sum_{l=1}^{L} \boldsymbol{A}^{(l)} \boldsymbol{X}_{t-l} + \epsilon_t, \tag{1}$$

where L is the maximum lag considered, $\boldsymbol{A}^{(l)}$ is a $k \times k$ matrix of coefficients and ϵ_t is a noise term with zero mean. In the linear regime, time series j does not Granger-cause series i if for all l $\boldsymbol{A}^{(l)}_{ij} = 0$. Tank et al. (2021) generalise the definition of Granger causality for nonlinear autoregressive models:

$$X_{ti} = g_i(X_{<t1}, \ldots, X_{<tk}) + \epsilon_{ti}, \tag{2}$$

where $X_{<ti} = (..., X_{(t-2)i}, X_{(t-1)i})$ denotes the history of time series i, and g_i is a nonlinear function mapping the lagged values of other k time series to series i. Granger non-causality is concluded between time series i and j if for all $(X_{<t1}, \ldots, X_{<tk})$ and all $X'_{<tj} \neq X_{<tj}$, $g_i(X_{<t1}, \ldots, X_{<tj}, \ldots, X_{tk}) = g_i(X_{<t1}, \ldots, X'_{<tj}, \ldots, X_{tk})$, implying that g_i does not depend on $X_{<tj}$.

In Tank et al. (2021) the function g_i is parameterised by a multilayer perceptron (MLP) regularised by group lasso penalties and trained with proximal gradient descent to shrink the input weights of lagged values of non-causal time series to zero. Bussmann et al. (2020) propose a neural additive VAR model with each time series expressed as a sum of nonlinear functions of the other time series. The nonlinear functions are parameterised by MLPs and the additive structure allows the contribution of each time series to be analysed separately.

Nauta et al. (2019) propose an attention based convolutional neural network with an explicit validation phase. The attention mechanism learns which time series are attended to during prediction, and interventions on potential causal time series are performed in the validation phase. Khanna and Tan (2020) infer Granger causal relations from a structured sparse estimate of internal parameters of statistical recurrent units (Oliva et al., 2017) trained for time series prediction.

A popular class of methods involves training two neural network time series prediction models and comparing their performances. One model would accept the past values of the target and exogenous variables as inputs, and the other accepts only the past target values. A statistically significant reduction in prediction error is a sign of Granger causality. In existing literature, these prediction models are often different variants of RNNs (Abbasvandi & Nasrabadi, 2019; Duggento et al., 2019; Wang et al., 2018) or MLPs (Orjuela-Canon et al., 2020). Our proposed approach falls within this class of methods however we argue that training two separate neural networks is inefficient and spurious conclusions could be reached due to differences in neural network hyperparameters. In the proposed architecture, a neural network with two simultaneously trained decoders is used to alleviate these issues.

With the exception of Nauta et al. (2019), all above-mentioned literature assumes causal sufficiency. How best to account for an unobserved confounder in Granger causal analysis is an open question. In Nauta et al. (2019), the model can only detect a latent confounder if it affects cause and effect with equal time lags. In this paper we consider a more challenging scenario involving different lags in the causal mechanisms. We follow a popular approach involving the use of neural networks to infer representations of the latent confounder (a substitute confounder). Louizos et al. (2017) propose a variational autoencoder to recover the joint distribution of the observed and latent variables which they use to estimate the average treatment effect (ATE) in a static setting. Trifunov et al. (2019) adapt the architecture in Louizos et al. (2017) to a time series setting for the estimation of ATE. Bica et al. (2020) propose a recurrent neural network architecture to build a factor model and estimate ATE using the inferred substitute confounders.

Outside of the deep learning domain, different methods can accommodate hidden confounders to different extents. Chu and Glymour

(2008) propose additive nonlinear time series model (ANLTSM) which can only deal with hidden confounders that are linear and instantaneous. Conditional independence based approaches LPCMCI (Gerhardus & Runge, 2020) and SVARFCI (Malinsky & Spirtes, 2018) detect hidden confounders by inferring a special edge type in the partial ancestral graph.

3. Methodology

Consider the following nonlinear autoregressive (NAR) model for time series i regressed on the histories of k other time series:

$$X_{ti} = g_i(X_{<t1}, \dots, X_{<tk}) + \epsilon_{ti}, \tag{3}$$

with nonlinear function g_i and white noise error term $\epsilon_{ti} \sim \mathcal{N}(0, \sigma_t^2)$. Since relationships between real world time series are often nonlinear, the definition of nonlinear Granger causality presented by Tank et al. (2021) is adopted in this study (see Section 2). More formally, time series j does not Granger cause series i if for all $(X_{<t1}, \dots, X_{<tk})$ and all $X'_{<tj} \neq X_{<tj}$, $g_i(X_{<t1}, \dots, X_{<tj}, \dots, X_{tk}) = g_i(X_{<t1}, \dots, X'_{<tj}, \dots, X_{tk})$. This implies that the prediction model g_i does not depend on the history of j $(X_{<tj})$, since substituting it with a different time series $(X'_{<tj})$ does not affect the prediction of X_{ti}. On the other hand, if series j does Granger cause series i, and j' does not, then the model with $X_{<tj}$ as input would lead to a lower prediction error of X_{ti} than if $X'_{<tj}$ was used as input instead: $(X_{ti} - g_i(X_{<t1}, \dots, X_{<tj}, \dots, X_{tk}))^2 < (X_{ti} - g_i(X_{<t1}, \dots, X'_{<tj}, \dots, X_{tk}))^2$.

The nonlinear function g_i can be modelled using a recurrent neural network as it can capture long range dependencies and complex temporal dynamics. The main challenge however is that the definition of Tank et al. (2021) assumes causal sufficiency (no confounding): all k time series are observed. In the presence of confounding, one observes only a subset of k: $(X_{<t1}, X_{<t2} \dots) \subseteq (X_{<t1}, \dots, X_{<tk})$; the use of traditional Granger causality tests in this case is known to be biased (Peters et al., 2017).

With access to proxy variables U, one can obtain representations of the latent confounder by approximating a function such that $\hat{Z} = f(U) \approx Z$; since f is likely to be a nonlinear function, neural networks could be used for this task. Note that $\hat{Z}$ and U do not need to have the same dimensions, since two proxies could result from the same confounder. Instead, the dimension of $\hat{Z}$ is a hyperparameter that is tuned during model selection. It is worth mentioning that since Granger causality only accounts for direct causal links (Eichler, 2013), one cannot simply use the proxies in a Granger causality test in place of the latent confounder (Louizos et al., 2017) since we see from Fig. 1 that there is no direct causal link between U and Y. Therefore, one must work backwards along the causal link $Z \to U$ to find a substitute confounder $\hat{Z}$ that can be used in place of Z.

Probabilistically, the output of the nonlinear autoregressive model given by (3) can be written as:

$$X_{ti} \sim \mathcal{N}(g_i(X_{<t1}, \dots, X_{<tk}), \sigma_t^2) = P(X_{ti} | X_{<t1}, \dots, X_{<tk}), \tag{4}$$

which is referred to as the predictive distribution of series i at time t conditioned on the histories of itself and other available time series. In this paper, neural networks are used to output the mean (g_i) and variance (σ_t^2).

Our proposed approach involves the use of multiple recurrent neural networks to parameterise predictive distributions. The full model predictive distribution is defined as $P(Y_{t+1} | Y_{1:t}, X_{1:t}, Z_{1:t})$. The restricted model distribution is defined as $P(Y_{t+1} | Y_{1:t}, Z_{1:t})$. Parameterising the two predictive distributions enables comparison of the predictive performances of two time series prediction models and a statistically significant reduction in prediction error from the restricted model to the full model is a sign of Granger causality $X \to Y$, or, more formally: $(Y_{t+1} - g(Y_{1:t}, X_{1:t}, Z_{1:t}))^2 < (Y_{t+1} - g(Y_{1:t}, Z_{1:t}))^2$.

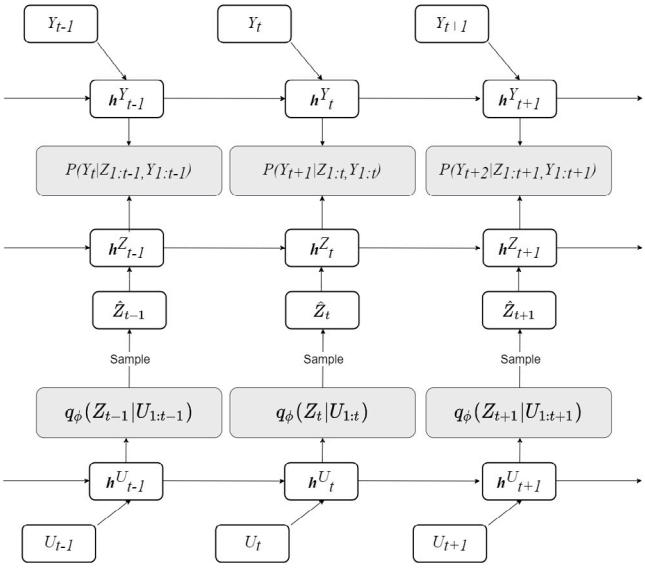

Fig. 2. Proposed architecture for the restricted model parameterised by multiple recurrent neural networks. q_ϕ is the inference distribution of Z conditioned on the proxy time series, from which samples of the substitute confounder $\hat{Z}$ can be obtained and used to parameterise the predictive distribution of Y.

To parameterise the full-model and restricted-model distributions, recurrent neural networks are used. The proposed model uses gated recurrent units (GRU) (Cho et al., 2014). The architecture of the restricted model is given in Fig. 2. Each GRU is characterised by a sequence of hidden states $\boldsymbol{h}_t^{(i)}$ which contains information of time series i up to time t. These hidden states are used as inputs to MLPs, which output the distribution parameters of the predictive and inference distributions. We propose to learn representations of the latent confounder Z using the available proxies U by parameterising the filtering distribution:

$$q_\phi(Z_t | U_{1:t}) = q_\phi(Z_t | \boldsymbol{h}_t^{(U)}). \tag{5}$$

The inferred representation $\hat{Z}_t$ of Z_t follows an isotropic Gaussian distribution:

$$\hat{Z}_t \sim q_\phi(Z_t | \boldsymbol{h}_t^{(U)}) = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{h}_t^{(U)}), \sigma^2(\boldsymbol{h}_t^{(U)})\boldsymbol{I}), \tag{6}$$

where the covariance matrix is diagonal. The dimension of $\hat{Z}_t$ is a tunable hyperparameter. The parameters of the filtering distribution are given by

$$(\boldsymbol{\mu}, \boldsymbol{\sigma}) = f_1(\boldsymbol{h}_t^{(U)}), \tag{7}$$

where f_1 is a mapping function approximated by an MLP. Let the hidden state $\boldsymbol{h}_t^{(U)} \in \mathbb{R}^{N_{h_U}}$ and $Z_t \in \mathbb{R}^{N_Z}$, the MLP takes as input a vector of size N_{h_U} and outputs a vector of size $N_Z \times 2$ (mean and variance). To ensure positivity of the standard deviation a softplus activation function is applied on the MLP output.

To avoid the need to train two separate time series prediction models, we propose the use of a dual-decoder setup. The restricted-model distribution is normal and given as

$$P(Y_{t+1} | Y_{1:t}, Z_{1:t}) = f_2(\boldsymbol{h}_t^{(Y)}, \boldsymbol{h}_t^{(Z)}). \tag{8}$$

The full-model distribution is also normal and expressed as

$$P(Y_{t+1} | Y_{1:t}, X_{1:t}, Z_{1:t}) = f_3(\hat{Y}_{t+1}^{res}, \boldsymbol{h}_t^{(X)}), \tag{9}$$

where $\hat{Y}_{t+1}^{res} \sim P(Y_{t+1} | Y_{1:t}, Z_{1:t})$ is the predicted value of Y_t from the restricted model and f_2 and f_3 are two MLP models which output the means and variances of the predictive distributions. The proposed dual-decoder setup is shown in Fig. 3 and $\hat{Y}_{t+1}^{full} \sim P(Y_{t+1} | Y_{1:t}, X_{1:t}, Z_{1:t})$. A combination of Figs. 2 and 3 represents the full architecture of the

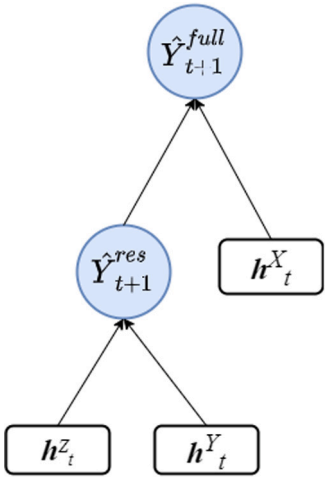

Fig. 3. Proposed dual-decoder setup where $\hat{Y}^{res}_{t+1}$ is a prediction sample drawn from the restricted-model distribution $P(Y_{t+1}|Y_{1:t}, Z_{1:t})$ shown in Fig. 2.

model where the output of the restricted-model serves as one of the inputs of the full-model.

For model optimisation the following objective function is maximised:

$$L = \sum_{t=1}^{T} \mathbb{E}_{\hat{Z}\sim q_\phi}[log P_{\theta_1}(Y_t|Y_{1:t-1}, X_{1:t-1}, Z_{1:t-1}) + log P_{\theta_2}(Y_t|Y_{1:t-1}, Z_{1:t-1})], \tag{10}$$

where the first and second terms correspond to the full and restricted model distributions respectively, and θ_1 and θ_2 are the model parameters to be optimised.

To infer the Granger causal relationship between X and Y in the presence of a latent confounder, we wish to check whether the inclusion of X in the full-model results in a statistically significant reduction in prediction error compared to the restricted model. With substitute confounders $\hat{Z}_{1:t}$ a two-sample t-test is performed to establish whether $Y_{t+1} \not\perp\!\!\!\perp X_{1:t}|\hat{Z}_{1:t}, Y_{1:t}$ (where $\perp\!\!\!\perp$ denotes independence); in such cases the conclusion X Granger-causes Y can be drawn, and vice versa. The mean-squared-error $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ was chosen as the error metric.

4. Experiments

We first demonstrate the model performance on two arbitrary synthetic datasets with known data generating processes. The nonlinear functions and noise levels have been set arbitrarily. The data generating processes for the two datasets are given by (11) and (12) respectively. In total, 1000 samples were generated, of which 800 were used for training, 100 for validation and 100 for testing.

4.1. Dataset 1

$$\begin{aligned} Z_t &= tanh(Z_{t-1}) + N(0, 0.01^2) \\ U_t &= Z_t^2 + N(0, 0.05^2) \\ X_t &= \sigma(Z_{t-2}) + N(0, 0.01^2) \\ No\ Granger &: Y_t = \sigma(Z_{t-4}) + N(0, 0.01^2) \\ Granger &: Y_t = \sigma(Z_{t-4}) + \sigma(X_{t-2}) + N(0, 0.01^2) \end{aligned} \tag{11}$$

where the hyperbolic tangent $tanh$ and sigmoid σ functions are used to introduce non-linearity into the system. The noise term is Gaussian of the form $N(\mu, std^2)$.

4.2. Dataset 2

The data generating processes for Z, U and X remain the same as in (11). The target series Y was generated using:

$$\begin{aligned} No\ Granger &: Y_t = Z_{t-3}Z_{t-4} + N(0, 0.5^2) \\ Granger &: Y_t = Z_{t-3}Z_{t-4} + X_{t-1}X_{t-2} + N(0, 0.5^2) \end{aligned} \tag{12}$$

4.3. River discharge dataset

To investigate the model performance on real-world time series, we use the river discharge dataset provided in Gerhardus and Runge (2020). This dataset describes the average daily discharges of rivers in the upper Danube basin. We consider measurements from the Iller at Kempten as X, the Danube at Dillingen as Y, and the Isar at Lenggries as the proxy variable. All three variables are potentially confounded by rainfall or other weather conditions (Gerhardus & Runge, 2020). The Iller discharges into the Danube within a day, implying an instantaneous causal link $X \to Y$. For the scope of Granger causality considered in this paper, the cause is required to precede the effect in time (Eichler, 2012) so we do not take into account instantaneous causal relationships. We therefore expect no Granger-causal relationship between X and Y. The dataset contains roughly 1000 entries, of which 80% were used for training, 10% for validation and 10% for testing.

4.4. Neural network parameters

The GRU hidden states $\boldsymbol{h}_t^{(X)}$, $\boldsymbol{h}_t^{(Y)}$ and $\boldsymbol{h}_t^{(Z)}$ have a dimension of 5, $\hat{Z}_t$ has a dimension of 1 for the synthetic datasets and 2 for the river discharge dataset, the MLPs f_1, f_2 and f_3 given in (7), (8), (9) respectively contain 1 hidden layer with 5 units for the synthetic datasets and 10 units for river discharge, a dropout rate of 0.3 and ReLU activation functions are chosen. The ADAM optimiser is used with a learning rate of 0.001. The neural networks were trained for 50 epochs. The sequence length used for model training is 20 with a batch size of 10. These parameters were selected using the validation set through random search.

4.5. Statistical testing

By comparing the sample prediction errors of the full and restricted models, we are able to infer whether a Granger causal relationship between X and Y exists. A two-sample t-test could be used as an additional verification step. For dataset 1&2 (Granger) consider the following null and alternative hypothesis:

$$\begin{aligned} H_0 &: \epsilon_{full} = \epsilon_{restricted} \\ H_1 &: \epsilon_{full} < \epsilon_{restricted}, \end{aligned} \tag{13}$$

where ϵ_{full} and $\epsilon_{restricted}$ are the mean prediction errors generated by the full and restricted models respectively. For dataset 1&2 (no Granger) consider the following alternative hypothesis:

$$H_1 : \epsilon_{full} > \epsilon_{restricted}. \tag{14}$$

The alternative hypothesis is chosen by comparing the sample mean prediction errors computed by the full and restricted models, I.e. the alternative hypothesis in (13) is chosen if the mean sample error of the full model is less than that of the restricted model, and vice versa. In cases where the mean sample errors of the two models differ significantly from one another, the statistical test is perhaps redundant. To perform the two-sample t-test, we generate $n = 50$ prediction samples from the restricted and full models and choose a significance level of $\alpha = 0.05$.

Table 1
Table showing the prediction errors of the full and restricted models, p values of two-sample t-tests and the inferred Granger causal relationship given by our model, LPCMCI and SVAR-FCI. The symbol × denotes that the model finds a Granger non-causal relationship between X and Y.

Dataset	Restricted-model error	Full-model error	p value	Ours	LPCMCI	SVAR-FCI
Dataset 1 (Granger)	$4.99 \times 10^{-2} \pm 6.00 \times 10^{-4}$	$1.76 \times 10^{-2} \pm 5.71 \times 10^{-5}$	<0.001	✓	✓	✓
Dataset 1 (no Granger)	$2.03 \times 10^{-2} \pm 4.00 \times 10^{-4}$	$3.54 \pm 2.00 \times 10^{-4}$	<0.001	×	✓	✓
Dataset 2 (Granger)	$2.07 \times 10^{-1} \pm 7.00 \times 10^{-4}$	$2.03 \times 10^{-1} \pm 9.75 \times 10^{-5}$	<0.001	✓	×	×
Dataset 2 (no Granger)	$1.56 \times 10^{-1} \pm 1.85 \times 10^{-4}$	$1.60 \times 10^{-1} \pm 5.73 \times 10^{-6}$	<0.001	×	×	×
River discharge	$4.85 \times 10^{-2} \pm 1.50 \times 10^{-3}$	$6.10 \times 10^{-2} \pm 1.12 \times 10^{-3}$	<0.001	×	×	×

Table 2
Sensitivity analysis of model performance with varying signal-to-noise ratio γ.

Dataset 1 (Granger) γ	*p value*	Dataset 2 (Granger) γ	*p value*
10.00	1.00	10.00	1.00
55.00	9.99×10^{-1}	21.25	9.99×10^{-1}
57.81	9.41×10^{-1}	26.88	5.44×10^{-2}
58.51	4.10×10^{-1}	**27.58**	4.73×10^{-3}
59.22	1.93×10^{-2}	28.28	$<1.00 \times 10^{-3}$
60.63	$<1.00 \times 10^{-3}$	29.69	$<1.00 \times 10^{-3}$
66.25	$<1.00 \times 10^{-3}$	32.50	$<1.00 \times 10^{-3}$
77.50	$<1.00 \times 10^{-3}$	55.00	$<1.00 \times 10^{-3}$
100	$<1.00 \times 10^{-3}$	100	$<1.00 \times 10^{-3}$

5. Results & discussion

In Table 1 we provide the prediction errors of the full and restricted models, the p values of the two-sample t-tests and the Granger causal relationship between X and Y inferred by our model, as well as those inferred by LPCMCI (Gerhardus & Runge, 2020) with $\alpha = 0.05$, maximum lag $L = 5$ and 4 preliminary iterations, and SVAR-FCI (Malinsky & Spirtes, 2018) with $\alpha = 0.05$ and $L = 5$. These are conditional independence based methods for inferring potential causal relationships and are capable of handling latent confounders.

It is evident from Table 1 that the $p\ value < 0.05$ for all the statistical tests. For dataset 1&2 (no Granger) and the river discharge dataset, we reject the null hypothesis that the mean prediction errors of the restricted and full models are equal and conclude that the inclusion of X to predict future values of Y results in a higher prediction error and therefore X does not Granger-cause Y. For dataset 1&2 (Granger) we reject the null hypothesis and conclude that the inclusion of X reduces the prediction errors of Y and therefore X Granger-causes Y. The proposed model correctly identifies the correct Granger-causal relationship in all scenarios, whereas LPCMCI and SVAR-FCI identify spurious relationships for dataset 1 (no Granger) and dataset 2 (Granger).

Real-world time series can be highly nonlinear and have different noise levels. The above analysis shows the proposed model is able to identify the Granger-causal relationship for various nonlinear functions and arbitrary noise levels. We investigate the robustness of our model by varying the signal-to-noise ratio defined as:

$$\gamma = \frac{\frac{1}{T}\sum_{t=1}^{T} |s_t|}{\sigma}, \tag{15}$$

where $|s_t|$ denotes the magnitude of the signal (Y_t without the noise term) at t and σ is the standard deviation of the noise term in the data generating process. For dataset 1&2 (Granger) we wish to find the critical γ below which the noise term becomes dominant and the model fails to identify the Granger-causal link between X and Y; to do this we vary the standard deviation σ of the noise term in (11) and (12). Starting with a rough range of $\gamma = 10$ to $\gamma = 100$, a bisection search strategy to find the critical value γ^*. A $p\ value < 0.05$ denotes Granger causality inferred by the proposed model. Results are shown in Table 2. For dataset 1 (Granger) we see that the critical value γ^* is approximately 59.22 (highlighted in bold), i.e. the Granger-causal link between X and Y for this set of stochastic time series can only be identified if $\gamma \geq 59.22$; for dataset 2(Granger) $\gamma^* \approx 27.58$.

Lastly, we tested the sensitivity of the model output to the sequence length $\tau \in \{4, 6, 8, 10, 12, 14, 16\}$ used in training for dataset 1&2 (Granger). We noted that all $p\ value < 0.001$, which suggests that our model is able to consistently identify the Granger-causal link given short and long τ used in training. This is desirable as it indicates that model results are not very sensitive to the choice of hyperparameters.

6. Conclusion

In this paper we have presented a deep-learning based approach to model nonlinear Granger-causality with in the presence of a latent confounder. Our model involves the use of multiple recurrent neural networks to parameterise a restricted-model distribution $P(Y_{t+1}|Y_{1:t}, Z_{1:t})$ and a full-model distribution $P(Y_{t+1}|Y_{1:t}, X_{1:t}, Z_{1:t})$. We generate prediction samples from the two distributions and we use a two-sample t-test to establish whether the inclusion of X helps to predict future values of Y given a learned representation of the confounder. To enable efficient comparison, we propose a dual-decoder setup, which avoids the need to train two separate models (as presented in many existing literature), and we believe this helps to reduce bias resulting from neural network hyperparameter tuning. We demonstrate the effectiveness of our model on both synthetic and real-world datasets, and we recognise that a high enough signal-to-noise ratio is required to correctly identify a Granger-causal link.

CRediT authorship contribution statement

Zexuan Yin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Paolo Barucca:** Methodology, Validation, Formal analysis, Investigation, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge Dr Fabio Caccioli and Dr Brooks Paige (both Dept of Computer Science, UCL) for their advice on the project and the manuscript.

References

Abbasvandi, Z., & Nasrabadi, A. M. (2019). A self-organized recurrent neural network for estimating the effective connectivity and its application to EEG data. *Computers in Biology and Medicine*, *110*(May), 93–107. http://dx.doi.org/10.1016/j.compbiomed.2019.05.012.

Bezruchko, B. P., Ponomarenko, V. I., Prokhorov, M. D., Smirnov, D. A., & Tass, P. A. (2008). Modeling nonlinear oscillatory systems and diagnostics of coupling between them using chaotic time series analysis: applications in neurophysiology. *Physics-Uspekhi*, *51*(3), 304–310. http://dx.doi.org/10.1070/pu2008v051n03abeh006494.

Bica, I., Alaa, A. M., & Van Der Schaar, M. (2020). Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *37th international conference on machine learning, ICML 2020, Vol. PartF16814* (pp. 861–872). arXiv:1902.00450.

Bussmann, B., Nys, J., & Latré, S. (2020). Neural additive vector autoregression models for causal discovery in time series data. arXiv preprint. URL: http://arxiv.org/abs/2010.09429. arXiv:2010.09429.

Chen, Y., Bressler, S. L., & Ding, M. (2006). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *Journal of Neuroscience Methods*, *150*(2), 228–237. http://dx.doi.org/10.1016/j.jneumeth.2005.06.011, arXiv:0608034.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Eighth workshop on syntax, semantics and structure in statistical translation (SSST-8)*. http://dx.doi.org/10.3115/v1/w14-4012, arXiv:1409.1259.

Chu, T., & Glymour, C. (2008). Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, *9*, 967–991.

De Brouwer, E., Arany, A., Simm, J., & Moreau, Y. (2020). Inferring causal dependencies between chaotic dynamical systems from sporadic time series. In *ICML workshop on the art of learning with missing values (Artemiss) 2020 Icml*.

Duggento, A., Guerrisi, M., & Toschi, N. (2019). Echo state network models for nonlinear granger causality. *BioRxiv*, 1–7. http://dx.doi.org/10.1101/651679.

Eichler, M. (2012). Causal inference in time series analysis. In *Causality: Wiley series in probability and statistics* (pp. 6–28). http://dx.doi.org/10.1002/9781119945710.ch22.

Eichler, M. (2013). Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of the Royal Society of London A (Mathematical and Physical Sciences)*, *371*(1997), http://dx.doi.org/10.1098/rsta.2011.0613.

Gerhardus, A., & Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders. In *34th conference on neural information processing systems NeurIPS*. URL: http://arxiv.org/abs/2007.01884. arXiv:2007.01884.

Granger, C. J. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438.

Hiemstra, C., & Jones, J. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, *49*(5), 1639–1664.

Khanna, S., & Tan, V. Y. F. (2020). Economy statistical recurrent units for inferring nonlinear granger causality. In *8th international conference on learning representations, ICLR 2020*. URL: http://arxiv.org/abs/1911.09879. arXiv:1911.09879.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, *2017-December*(Nips), 6447–6457, arXiv:1705.08821.

Malinsky, D., & Spirtes, P. (2018). Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal disocvery 2010* (pp. 1–25).

Marcinkevičs, R., & Vogt, J. E. (2021). Interpretable models for granger causality using self-explaining neural networks. In *9th international conference on learning representations, ICLR 2021*. URL: http://arxiv.org/abs/2101.07600. arXiv:2101.07600.

Marinazzo, D., Liao, W., Chen, H., & Stramaglia, S. (2011). Nonlinear connectivity by Granger causality. *NeuroImage*, *58*(2), 330–338. http://dx.doi.org/10.1016/j.neuroimage.2010.01.099.

Moraffah, R., Sheth, P., Karami, M., Bhattacharya, A., Wang, Q., Tahir, A., Raglin, A., & Liu, H. (2021). Causal inference for time series analysis: Problems, methods and evaluation. arXiv preprint. URL: http://arxiv.org/abs/2102.05829. arXiv:2102.05829.

Nauta, M., Bucur, D., & Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, *1*(1), 312–340. http://dx.doi.org/10.3390/make1010019.

Oliva, J. B., Poczos, B., & Schneider, J. (2017). The statistical recurrent unit. In *34th international conference on machine learning, ICML 2017, Vol. 6* (pp. 4098–4107). arXiv:1703.00381.

Orjuela-Canon, A. D., Freund, J. A., Jutinico, A., & Cerquera, A. (2020). Granger causality analysis based on neural networks architectures for bivariate cases. In *Proceedings of the international joint conference on neural networks* (pp. 1–6). http://dx.doi.org/10.1109/IJCNN48605.2020.9206977.

Pearl, J. (2010). On measurement bias in causal inference. In *Proceedings of the 26th conference on uncertainty in artificial intelligence, UAI 2010* (pp. 425–432). arXiv:1203.3504.

Peters, J., Janzing, D., & Schlkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms* (pp. 203–210). The MIT Press.

Rahimi, M., Davoodi, R., & Moradi, M. H. (2020). Deep fuzzy model for nonlinear effective connectivity estimation in the intuition of consciousness correlates. *Biomedical Signal Processing and Control*, *57*, Article 101732. http://dx.doi.org/10.1016/j.bspc.2019.101732.

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, *3*(1), http://dx.doi.org/10.1186/s40535-016-0018-x.

Stips, A., MacIas, D., Coughlan, C., Garcia-Gorriz, E., & Liang, X. S. (2016). On the causal structure between CO2 and global temperature. *Scientific Reports*, *6*(February), 1–9. http://dx.doi.org/10.1038/srep21691.

Tank, A., Covert, I., Foti, N., Shojaie, A., & Fox, E. B. (2021). Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14. http://dx.doi.org/10.1109/TPAMI.2021.3065601, arXiv:1802.05842.

Trifunov, V. T., Shadaydeh, M., Runge, J., Eyring, V., Reichstein, M., & Denzler, J. (2019). Nonlinear causal link estimation under hidden confounding with an application to time series anomaly detection. In *LNCS*: *vol. 11824*, *41st DAGM German conference on pattern recognition, DAGM GCPR 2019* (pp. 261–273). http://dx.doi.org/10.1007/978-3-030-33676-9_18.

Vicente, R., Wibral, M., Lindner, M., & Pipa, G. (2011). Transfer entropy-a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, *30*(1), 45–67. http://dx.doi.org/10.1007/s10827-010-0262-3.

Wang, Y., Lin, K., Qi, Y., Lian, Q., Feng, S., Wu, Z., & Pan, G. (2018). Estimating brain connectivity with varying-length time lags using a recurrent neural network. *IEEE Transactions on Biomedical Engineering*, *65*(9), 1953–1963. http://dx.doi.org/10.1109/TBME.2018.2842769.

Yuan, A. E., & Shou, W. (2020). Data-driven causal analysis of observational time series: A synthesis. *BioRxiv*, URL: https://doi.org/10.1101/2020.08.03.233692.