RELATION DISCOVERY IN NONLINEARLY RELATED LARGE-SCALE SETTINGS

Ali Vosoughi[⋆]*, IEEE Member, Adora DSouza*[⋆]*, Anas Abidin*[‡]*, and Axel Wismüller*[⋆,†,‡]

[⋆] Dept. of Elec. and Comp. Eng.,[†] Dept. of Imag. Sc.,[‡] Dept. of Biomed. Eng.,
University of Rochester, New York, USA, 14623, [⋄] Inst. of Clinical Radiology,
Ludwig Maximilian University, Munich, Germany

ABSTRACT

Causal inquiries provide crucial insight into the advancement of scientific discoveries. In real-world studies like climatology, sensory data acquired from nodal measurements are nonlinearly related and complex. At the same time, they have information from millions of sensors with only a few decades' temporal samples, which leads to the curse of dimensionality in large-scale systems. Despite a rich literature on causal discovery, the problem is challenging for large-scale datasets. We put forth a novel method that utilizes a radial basis function (RBF) to tackle curse-of-dimensionality in complex systems. The proposed method is probabilistic, encompasses nonlinear relations, and is suitable for large-scale data in two steps. Extensive simulations on synthetic data of different sizes and real-world climatology data show that our method outperforms all other methods when nodal observations are temporally scarce.

Index Terms— relation discovery, causal learning, network inference, causality, nonlinear relationships

1. INTRODUCTION

Causal representations are meaningful and apply to many real-world problems, evidenced by DARPA's machine common sense project that aim to reach the human-level cognitive abilities [1]. Not surprisingly, in the recent decade, causality gained the most traction to make generalizable models with adequate robustness, explainability, and the ability to learn semantically meaningful representations [2]. It was development of *do-calculus* and completeness of identifiability by *Pearl* that has paved the way for further developments in the field of causality [3]. However, in the absence of a structural graph, the use of *do-calculus* remains unsettled [4] and inference methods are required to obtain network topology from the observed data. Information-theoretic methods such as Granger causality (GC) are widely adopted for causal inference from data given the faithfulness assumption [4].

It was *Wiener* who developed the autoregressive model [5] in 1956, and *Granger* utilized it to define causality in terms of prediction quality [6]. GC initially stated for linear relations, and it states that a time-series A is a cause for time-series B if we can better predict B *"given the past of all information"* than *"given the past of all information in the universe"* excluding A [6]. Later, a wide variety of literature on the causal discovery of nonlinear relations from nodal time-series data in Granger's framework has been proposed [7, 8, 9, 10, 11, 12, 13, 14, 15], however, it is yet an open problem mainly due to the curse-of-dimensionality and ill-posedness besides vulnerability to confounding variables [16, 17]. Note that the real-world studies such as climatology involve millions of sensors, while their observations may be limited by time to only a few decades of recording [18], in which case the inverse problems quickly become ill-posed, exacerbated when redundant variables are present.

In [19] *Ancona* presents conditions for nonlinear inference of causality, and *Marinazzo* [20] develops a method called kernel Granger causality that leverages kernel transformation, while [19] is a bivariate method that cannot capture confounding variables, and [20] is unsuited for large-scale systems due to over-fitting. In parallel a method based on transfer entropy (TE) by *Schreiber* [10], and mutual information (MI) by k-nearest neighborhood (KNN) estimation of density by *Kraskov* [7] were proposed, which form the information-theoretic perspectives to causality inference from the time-series. Unfortunately, TE and MI-based techniques are computationally expensive in large-scale problems and infamous due to their dependency on estimating the probability densities. We provide an extensive analysis of recent algorithms to underscore the drastic increase in the computation of large-scale problems. Different methods based on inducing sparsity by regularization theory [21] are proposed, which their inclusive version can be seen in elastic nets (EN) Granger causality [12]; however, the models in regularization-based causality estimation are linear. Our proposed method differs from the above methods for two reasons. First, since it is a nonlinear GC method, it is equivalent to the TE method for Gaussian variables [22], while it is also suitable for large-scale problems. Second, the method we present allows us to utilize short time series on a large scale. We believe our contribution provides immense capabilities to the existing literature, affirmed throughout extensive simulations on synthetic and real-world datasets.

The rest of the paper is as follows. In Section 2, we provide preliminaries for causality and provide assumptions in which we combine it with signal processing. The proposed method is given in Section 3, supported by extensive simulations and conclusions in the consequent sections.

2. PRELIMINARIES

We use i, j and Y_i, Y_j in abuse of notions to point to the nodes and random variables over the nodes, scalars are denoted by small letters, vectors are shown in small bold letters, and matrices are in

This research was partially funded by the American College of Radiology (ACR) Innovation Award "AI-PROBE: A Novel Prospective Randomized Clinical Trial Approach for Investigating the Clinical Usefulness of Artificial Intelligence in Radiology" (PI: Axel Wismüller) and an Ernest J. Del Monte Institute for Neuroscience Award from the Harry T. Mangurian Jr. Foundation (PI: Axel Wismüller). This work was conducted as a Practice Quality Improvement (PQI) project related to American Board of Radiology (ABR) Maintenance of Certificate (MOC) for A.W. This work has been partially supported by the National Science Foundation (NSF) under Grant DGE-1922591.

Code is available: https://github.com/ali-vosoughi/CausalClimate

bold capital, and calligraphic letters $\mathcal{N}, \mathcal{E}$, and $\mathcal{Y}$ represent the set of nodes, set of weighted edges, and input space, correspondingly.

Causality assumptions: Consider a *directional graphical model* $G = (\mathcal{N}, \mathcal{E})$ consisting of the set of nodes $\mathcal{N} = \{i | i = 1, \ldots, N\}$ and the set of directed edges $\mathcal{E}$, whose topology is unknown, but time-series $\{y_{it}\}_{t=1}^{T}$ are observed per node i, over T time intervals. The *edges are causal*, meaning that every parent i is a direct cause of all its children $\mathcal{C}(i)$, and therefore i is a parent of j, denoting as $i \in \mathcal{P}(j)$. Furthermore, we assume that measurements $\{y_{it}\}_{t=1}^{T}$ are *causally sufficient*, since we cannot test unconfoundedness, and cannot guarantee that it is satisfied [23]. The ultimate goal is to estimate causal quantities δ_{ij} between each pair of the nodes i and j in the presence of all other nodes $Z = \mathcal{N} \setminus \{i, j\}$, and derive the topology of the graph as $\mathcal{E} = \{\delta_{ij} | i, j \in \mathcal{N}\}$, such that $i \in \mathcal{P}(j)$ if $\delta_{ij} > 0$ else $i \notin \mathcal{P}(j)$. We assume that δ_{ij} is *identifiable*, , which requires to have four assumptions: unconfoundness, positivity, consistency, and no interference [3]. Therefore, we can infer causality from purely statistical measures. We use *faithfulness* assumption to obtain causal relations among the nodes $\mathcal{N}$, which allows us to infer *d-separations* in the graph from in dependencies in the distributions $Y_i \perp\!\!\!\perp_G Y_j | Z \Leftarrow Y_i \perp\!\!\!\perp_P Y_j | Z$. The notion $\perp\!\!\!\perp$ implies independence in the graph G, or the distribution $P(Y_1, \ldots, Y_N)$ where Y_i denotes the random variable corresponding to node i. Note that faithfulness is a weak assumption [4]; however, without it, we cannot infer the topology of the causal graph from the observational time-series. The reason is related to global Markov assumption, where given that P is Markov with respect to graph G one can use G to infer independencies in P. However, in case of the unstructured time-series data, the graph topology is priory unknown and we cannot use Markov assumption before obtaining the graph topology.

Causal quantities: Vector autoregressive model is widely adopted framework to infer Granger causality in multivariate time-series, in which a *Granger causality index* δ_{ij} can be defined as a quantification of prediction quality as follows: Each observed $\mathbf{y}_t := [y_{1t}, \ldots, y_{Nt}]^\top$ is a linear combination of the time-lagged versions of the measurements $\{\{y_{i(t-\ell)}\}_{i=1}^{N}\}_{\ell=1}^{L}$. Let $\mathbf{A}^\ell \in \mathbb{R}^{N \times N}$ denote the time-lagged parameters matrix, with $[\mathbf{A}^\ell]_{ij} = a_{ij}^\ell$, and a_{ij}^ℓ as model coefficients over a lag of ℓ time points. Given the time-lagged multivariate time-series $\{\mathbf{y}_{(t-\ell)}\}_{t=\ell}^{L}$, where the goal is to estimate the model parameter matrices $\{\mathbf{A}^\ell\}_{\ell=1}^{L}$ in:

$$\mathbf{y}_t = \sum_{\ell=1}^{L} \mathbf{A}^\ell \mathbf{y}_{(t-\ell)} + \mathbf{e}_t, \tag{1}$$

and minimize the residual errors $\mathbf{e}_t$ using an optimization method, such as ordinary least squares (OLS), the lasso, the ridge, or the elastic-net regressions [24]. Let $\mathbf{E} := [\mathbf{e}_\ell, \ldots, \mathbf{e}_T]$ be the matrix of residuals of (1) for the full system including all nodes $\mathcal{N}$, and let $\mathbf{E}^{i-} := [\mathbf{e}_\ell^{i-}, \ldots, \mathbf{e}_T^{i-}]$ to be the same matrix when the node i is excluded $\mathcal{N} \setminus \{i\}$ from the VARM, where each $\mathbf{e}_t^{i-}$ is obtained using $\mathbf{y}_t^{i-} := [y_{1t}, \ldots, y_{(i-1)t}, y_{(i+1)t}, \ldots, y_{Nt}]^\top$ in expression (1). Then, the derivation of error covariance matrices $\Sigma = cov(\mathbf{E}, \mathbf{E})$ and $\Sigma^{i-} = cov(\mathbf{E}^{i-}, \mathbf{E}^{i-})$ will be straightforward. Based on Σ of the full VARM and Σ^{i-} of the VARM without i, the degree of information flow from node i to node j can be quantified by $\ln(\Sigma_j^{i-}/\Sigma_j)$, where Σ_j^{i-} and Σ_j denote the diagonal entries of Σ^{i-} and Σ associated to node j, respectively. Consequently, the topology of the causal graph can be inferred using:

$$\delta_{ij} = \begin{cases} \ln(\Sigma_j^{i-}/\Sigma_j), & \ln(\Sigma_j^{i-}/\Sigma_j) > 0 \\ 0, & \ln(\Sigma_j^{i-}/\Sigma_j) \leq 0, \end{cases} \tag{2}$$

as $\mathcal{E} = \{\delta_{ij} | i, j \in \mathcal{N}, i \in \mathcal{P}(j)$ if $\delta_{ij} > 0$ else $i \notin \mathcal{P}(j)\}$. Given nodal measurements $\{y_{it}\}_{t=1}^{T}$ per node $i = 1, \ldots, N$, the aim is to obtain causal quantities by nonlinear transformation, preserving nonlinear relations and solving the curse-of-dimensionality.

3. METHOD

The linear model in (1) cannot capture nonlinear relations, therefore to preserve nonlinear relations and solve the curse of dimensionality, we proceed as follows. Let $\mathbf{x}_t^j := \{y_{j(t-\ell)}\}_{\ell=1}^{T}$ for t from $L+1$ to T be a column vector $\mathbf{x}_t^j \in \mathbb{R}^L$. Using *Taken's theorem* [25] we define the state space representations as $\mathcal{D} := \{\mathbf{x}_t, y_t\}_{t=L+1}^{T}$, where $\mathbf{x}_t = \{\mathbf{x}_t^j\}_{j \in \mathcal{N}}$, and $\mathbf{x}_t^{i-} = \{\mathbf{x}_t^j\}_{j \in \mathcal{N} \setminus i}$, such that $\mathbf{x}_t \in \mathbb{R}^{NL}$ and $\mathbf{x}_t^{i-} \in \mathbb{R}^{(N-1)L}$ are column vectors. The embedding dimension L can be chosen using Cao's method [26]. The stability of Cao's method is less vulnerable to the number of temporal samples that are available from the datasets, which is the main point of using Cao's method as compared to other Takens-type embedding methods [26]. The aim is to obtain the influence of source time series i on all other time series $j \in \mathcal{N} \setminus i$. The nonlinear equivalent of (1) is given by [27] as $\mathbf{y}_t = \sum_{j=1}^{N} \alpha_j \phi_j(\mathbf{x}_t^j) + \mathbf{e}_t$, which also can be written as:

$$\mathbf{y}_t = \sum_{j \neq i} \alpha_j \phi_j(\mathbf{x}_t^j) + \alpha_i \phi_i(\mathbf{x}_t^i) + \mathbf{e}_t \tag{3}$$

To solve the curse-of-dimensionality while encompassing nonlinear relationships, we adopt the RBF neural network in combination of partition function as the nonlinear transformation $\sum_{j \neq i} \alpha_j \phi_j(\mathbf{x}_t^j) = W_f f(\mathbf{x}_t^{i-})$ and $\alpha_i \phi_i(\mathbf{x}_t^i) = W_g g(\mathbf{x}_t^i)$, which satisfies the independence and identically distributed ($i.i.d.$) condition for f and g for causality inference in the Granger's setting [20]. As shown in Fig. 1, the aim is to set up two different radial basis functions f and g to infer the relations between the nodal observations. These two f and g are activation functions of the radial basis function, and W_f, W_g are the weights that can be estimated by minimizing the prediction error. Such an architecture will require one to cluster $N \times N$ times to achieve cluster centers for both g and f, however, here to alleviate computational cost we obtain a global cluster centers for f, therefore reducing computational cost to only $1 \times N$. To do so, we obtain k_g cluster centers for the each $\mathbf{x}_t^i$ using k-means algorithm and obtain cluster centers $V^\top \in \mathbb{R}^{k_g \times L}$. Activation function $g(\mathbf{x}_t^i) = \{P_{g_j}(\mathbf{x}_t^i)\}_{j=1}^{k_g}$ is calculated as:

$$P_{g_j}(\mathbf{x}_t^i) = \frac{e^{-||\mathbf{x}_t^i - \mathbf{v}_j||^2/\sigma^2}}{\sum_{k=1}^{k_g} e^{-||\mathbf{x}_t^i - \mathbf{v}_k||^2/\sigma^2}}, \tag{4}$$

where, $j \in \{1, ..., k_g\}$ and $P_{g_j}(\mathbf{x}_t^i)$ is the probability that i-th node is associated with j-th cluster. This is the difference between radial basis function, and generalized radial basis function, as the use of partition function at the denominator of (4) assures that the the sum over all clusters k_g sums up to 1. The parameter σ is the scaling factor of the Gaussian kernel function and can be obtained using methods such as automatic relevance determination [28].

Analogously, cluster centers $U^\top \in \mathbb{R}^{k_f \times NL}$ are calculated for the state space $\mathbf{x}_t \in \mathbb{R}^{NL}$, where k_f is the number of clusters obtained with k-means clustering. Activation function $f(\mathbf{x}_t^{i-}) = \{P_{f_j}(\mathbf{x}_t^{i-})\}_{j=1}^{k_f}$ is as $P_{f_j}(\mathbf{x}_t^{i-}) = \frac{e^{-||\mathbf{x}_t^{i-} - \mathbf{u}_j||^2/\sigma^2}}{\sum_{k=1}^{k_f} e^{-||\mathbf{x}_t^{i-} - \mathbf{u}_j||^2/\sigma^2}}$, where, $j \in \{1, \ldots, k_f\}$. Then by rewriting the equation (3) in the following form $\mathbf{e}_t = \min_{W_f^*, W_g^*} \|\mathbf{y}_t - (W_f f(\mathbf{x}_t^{i-}) + W_g g(\mathbf{x}_t^i))\|$,

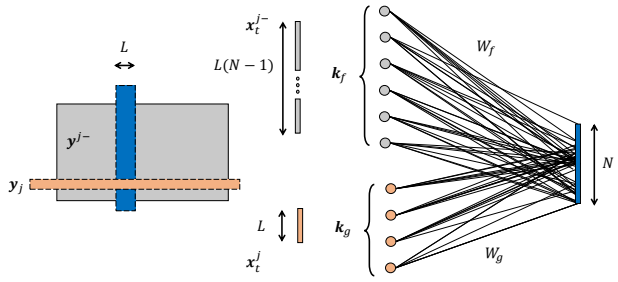

Fig. 1: The aim is to set up two different radial basis functions $f : \mathbb{R}^{(N-1)d} \mapsto \mathbb{R}^{k_f}$ and $g : \mathbb{R}^d \mapsto \mathbb{R}^{k_g}$ that help us to infer the relations between the nodal observations. These two f and g are activation functions of the radial basis function, and $W_f \in \mathbb{R}^{N \times k_f}$, and $W_g \in \mathbb{R}^{N \times k_g}$ are the weights that can be estimated by minimizing the prediction error. Note that k_f cluster centers remain unchanged throughout the process to reduce computational cost.

where $f : \mathbb{R}^{(N-1)d} \mapsto \mathbb{R}^{k_f}$, $g : \mathbb{R}^d \mapsto \mathbb{R}^{k_g}$, $W_f \in \mathbb{R}^{N \times k_f}$, and $W_g \in \mathbb{R}^{N \times k_g}$, and comparing it to the when the data of i-th node is removed as $\mathbf{e}_t^{i-} = \min_{W_f^{*\prime}} \|\mathbf{y}_t - W_f' f(\mathbf{x}_t^{i-})\|$, the topology will be obtained as discussed in Section 2. The complete algorithm is shown in Algorithm 1.

Algorithm 1 Algorithm for the proposed method

Require: $L, \mathbf{Y}, k_f, k_g$

Ensure: $\mathcal{E} \leftarrow \{\ln(\frac{\mathbf{\Sigma}_j^{i-}}{\mathbf{\Sigma}_j})\}_{j=1}^N$

k_f cluster centers $\mathbf{U} \leftarrow \{\mathbf{x}_t, \mathbf{y}_t\} \leftarrow \mathbf{Y} \leftarrow \text{normalize}(\mathbf{Y})$

for $j = 1$ to N $(\forall t)$ **do**

k_g cluster centers $\mathbf{V} \leftarrow \{\mathbf{x}_t^j, \mathbf{x}_t^{j-}\} \leftarrow \mathbf{x}_t$

$\mathbf{e}_t \xleftarrow{f(.),g(.)} \mathbf{x}_t^{j-}$, and $\mathbf{e}_t^{j-} \xleftarrow{f(.)} \mathbf{x}_t^{j-}$

$\mathbf{\Sigma}_j, \mathbf{\Sigma}_j^{j-} \leftarrow cov(\mathbf{e}_t), cov(\mathbf{e}_t^{j-})$

end for

4. NUMERICAL TESTS

Extensive simulations on a benchmark of the synthetic datasets with known ground-truth and a real-world dataset in a downstream task are presented in this section.

Tests on synthetic datasets: We used five various networks with various topologies, each with 50 random realizations, to test our method and some of the state-of-the-art methods, namely mutual information using Kraskov estimation (2018) [8], transfer entropy (TE) using Kraskov method (2018) [7, 8], and multivariate GC with elastic net regularization 2020 [21, 12], and multivariate Granger causality 2020 [29, 12] as the benchmark. The topology of the datasets are shown in Fig. 2. We used two 3-nodes with v-structure and immorality topologies, and one 5-nodes dataset with nonlinear dependencies, and two 34-node datasets with complex relations that are famous as Zachary club networks [30].

We used datasets to address two challenges: 1) Fig. 3 shows the case of which datasets have enough time samples for different topologies, and 2) Fig. 4 shows when the trend of different algorithms by reducing the number of time samples. AUROC is used to compare the performance in the box plots. It is clearly seen in Fig. 3 and 4 that our algorithm is more robust to the size of the network, nonlinearity in dependencies, and *short time-series*. Note that only our algorithm

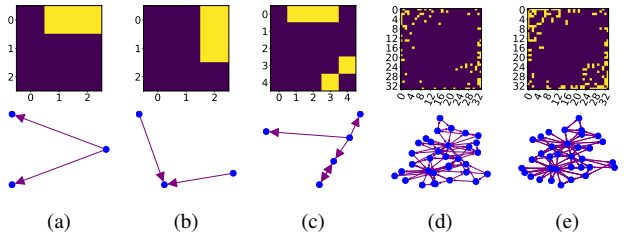

Fig. 2: The topologies of datasets, a) V-structure, b) Immortality, c) 5-Node nonlinear, d) Zachary 1, e) Zachary 1, which have a different number of nodes $\{3, 3, 5, 34, 34\}$, and different linear and nonlinear characteristics. Dataset equations are detailed in [11, 20, 30].

Table 1: Simulation time (in seconds) for T=500 of synthetic datasets in equal condition. Datasets are listed as V-structure: 3-Vs., 3-Immortality: 3-Im., 5-node nonlinear: 5-Nd., 34-node nonlinear (1): 34-Z1 , 34-node nonlinear (2): 34-Z2.

	MI	TE	EN	GC	Ours
3-Vs.	826	658	99	**0.7**	45
3-Im.	849	534	73	**0.7**	44
5-Nd.	1391	1191	387	**1.2**	176
34-Z1	135513	76818	19440	3576	**749**
34-Z2	158374	92623	26154	1580	**788**

is proposed for large-scale problems, we show how our algorithm works with the size of network in Table 1.

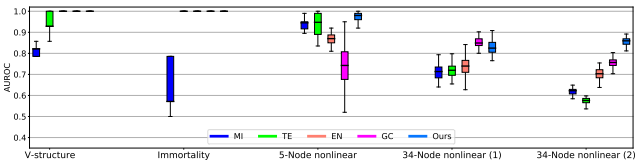

Fig. 3: Simulations of various algorithms on different datasets for T=500 are shown. Each column shows a different dataset (see titles). From left to right: v-structure, immortality, 5-node nonlinear, 34-node nonlinear (1), and 34-node nonlinear (2). Five different methods (MI, TE, EN, GC, and the proposed method) are tested to compare the performance. Each dataset has 50 different replications. The confidence intervals of 95 percent are shown as shaded areas around the means. Our proposed method (light blue, denoted by *Ours*) significantly outperforms the counterparts for all network sizes.

Tests on real datasets: We examine our method on climatology data of average daily discharges of rivers in the upper Danube basin by three stations located on the Iller at Kempten (IK), the Danube at Dillingen (DD), and the Isar at Lenggries (IL). The data are available through Bavarian Environmental Agency at *https://www.gkd.bayern.de*, and we use the measurements of three years (2017-2019). As shown in Fig. 5, there is a causal relationship IK $\rightarrow$ DD since IK discharges into the DD upstream after a day, while there is no causal relationship between IL$\rightarrow$IK and IL$\rightarrow$DD (and *vice versa*). Statistical analyses are vulnerable to detecting spurious connections due to confounder variables such as rainfall or other weather conditions. We emphasize that our proposed method detects a causal relationship between IK and DD while detecting no relationships between IL-IK and IL-DD, accurately unconfounding

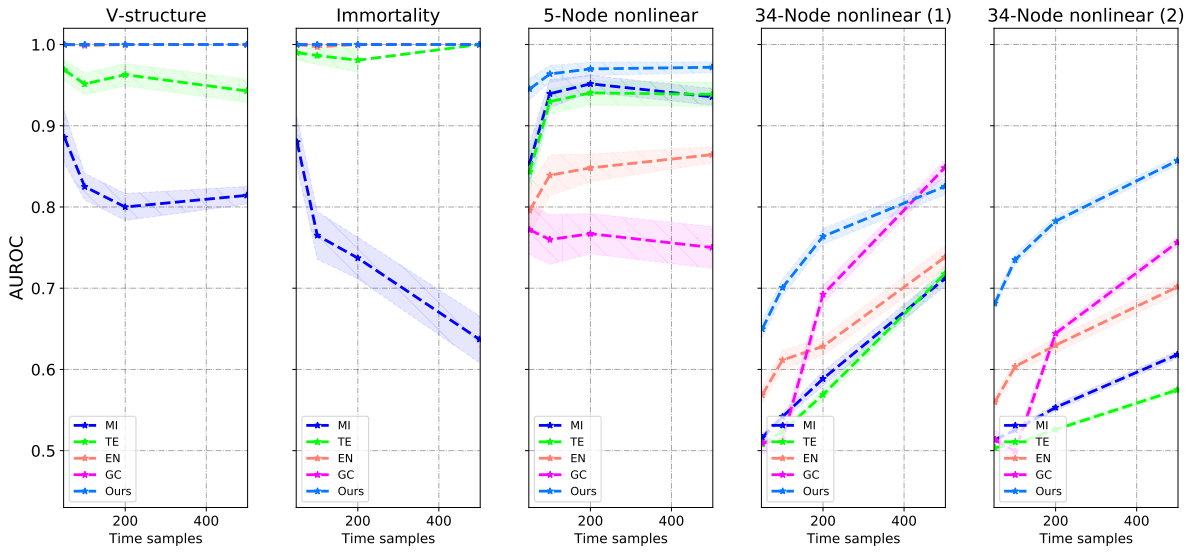

Fig. 4: Simulations of various algorithms on different datasets are shown. Each column shows a different dataset (see titles). From left to right: V-structure, immortality, 5-node nonlinear, 34-node nonlinear (1), and 34- node nonlinear (2). Five different methods (MI, TE, EN, GC, and the proposed method) are tested for various time samples T= {50, 100, 200, 500} to compare the performance on the short time-series. Each dataset has 50 different replications. The confidence intervals of 95 percent are shown as shaded areas around the means. The time samples are increasing from left to right. Our proposed method (light blue, denoted by *Ours*) is significantly more robust than the counterparts to the short time series and for all topologies.

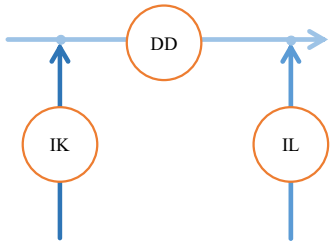

Fig. 5: The average daily discharges of rivers in the upper Danube basin, the Iller at Kempten (IK), the Danube at Dillingen (DD), and the Isar at Lenggries (IL) are nonlinearly related. Only IK causes DD, which is correctly detected by the proposed method.

spurious variables. Contrastingly, the Kraskov's TE [8, 10] wrongly detects the spurious connections as the DD→IK, DD→IL, IL→IK, and IL→DD altogether, for the best parameters. The analysis of rivers' underlying dynamics agrees to expert knowledge [31], which shows the capacity of our to distinguish between confounding and causal variables in a real-world pattern.

5. CONCLUSIONS

This paper puts forth a novel method to discover nonlinear relations from short temporal observations. The method aims to solve the curse of dimensionality in large-scale systems, incorporating nonlinear relationships between nodes. We propose a novel bipartite architecture by leveraging the radial basis function that uses partition function to probabilistically relate each variable's dependency to a cluster and provides a framework for Granger causality inference using radial basis. The information-theoretic approach that we provide is equivalent to transfer entropy for Gaussian variables, advancing the existing nonlinear methods to large-scale problems and drastically reducing the computational cost associated with existing methods. Our method's advantage rests in its large-scale approach that encompasses nonlinear relations for short time series. Extensive simulations on synthetic and real-world datasets affirm the superiority of our method compared to all competing state-of-the-art causality inference methods.

6. REFERENCES

[1] Matt Turek, "Machine Common Sense," February 16, 2022.

[2] Bernhard Schölkopf et al., "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, February 2021.

[3] Ilya Shpitser et al., "Complete Identification Methods for the Causal Hierarchy," *Journal of Machine Learning Research*, vol. 9, pp. 1941–1979, September 2008.

[4] Jonas Peters et al., *Elements of causal inference: foundations and learning algorithms*, The MIT Press, 2017.

[5] Norbert Wiener, "The theory of prediction," *Modern mathematics for engineers*, 1956.

[6] Clive WJ Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, August 1969.

[7] Alexander Kraskov et al., "Estimating Mutual Information," *Physical Review E*, vol. 69, no. 6, pp. 1–16, June 2004.

[8] Leonardo Novelli et al., "Large-Scale Directed Network Inference with Multivariate Transfer Entropy and Hierarchical Statistical Testing," *MIT Press*, vol. 3, no. 3, pp. 827–847, July 2019.

[9] Jakob Runge et al., "Inferring Causation from Time Series in Earth System Sciences," *Nature Communications*, vol. 10, no. 1, pp. 1–13, June 2019.

[10] Thomas Schreiber, "Measuring Information Transfer," *Physical Review Letters*, vol. 85, no. 2, pp. 1–4, July 2000.

[11] Daniele Marinazzo et al., "Nonlinear Connectivity by Granger Causality," *Neuroimage*, vol. 58, no. 2, pp. 330–338, September 2011.

[12] T. Wu et al., "Discovering nonlinear relations with minimum predictive information regularization," *arXiv preprint arXiv:2001.01885*, January 2020.

[13] Yanning Shen et al., "Nonlinear structural vector autoregressive models with application to directed brain networks," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5325–5339, September 2019.

[14] Gabriel Schamberg et al., "Measuring sample path causal influences with relative entropy," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2777–2798, 2019.

[15] Rongjin Ma et al., "Causality analysis based on matrix transfer entropy," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, September 2018, pp. 1–6.

[16] Jakob Runge et al., "Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets," *Science Advances*, vol. 5, no. 11, pp. 1–15, November 2019.

[17] Yuri Antonacci et al., "Testing Different Methodologies for Granger Causality Estimation: A Simulation Study," in *European Signal Processing Conference*, January 2021, pp. 940–944.

[18] Gustau Camps-Valls et al., "A Perspective on Gaussian Processes for Earth Observation," *National Science Review*, vol. 6, no. 4, pp. 616–618, July 2019.

[19] Nicola Ancona et al., "Radial Basis Function Approach to Nonlinear Granger Causality of Time Series," *Physical Review E*, vol. 70, no. 5, pp. 1–7, November 2004.

[20] Daniele Marinazzo et al., "Kernel-Granger Causality and the Analysis of Dynamical Networks," *Physical review E*, vol. 77, no. 5, pp. 1–9, May 2008.

[21] Hui Zou et al., "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, April 2005.

[22] Lionel Barnett, Adam B Barrett, and Anil Seth, "Granger Causality and Transfer Entropy are Equivalent for Gaussian Variables," *Physical Review Letters*, vol. 103, no. 23, pp. 1–4, December 2009.

[23] Charles Manski, *Partial Identification of Probability Distributions*, Springer Science & Business Media, 2003.

[24] J. Friedman et al., "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1, August 2010.

[25] Floris Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*, pp. 366–381. Springer, 1981.

[26] Liangyue Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D: Nonlinear Phenomena*, vol. 110, no. 1-2, pp. 43–50, December 1997.

[27] N. Lim et al., "Operator-valued kernel-based vector autoregressive models for network inference," *Machine learning*, vol. 99, no. 3, pp. 489–513, June 2015.

[28] Carl Edward Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*. Springer, 2003, pp. 63–71.

[29] Clive Granger, "Some Recent Development in a Concept of Causality," *Journal of Econometrics*, vol. 39, no. 1-2, pp. 199–211, September 1988.

[30] Wayne W Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, December 1977.

[31] Linda Mhalla et al., "Causal Mechanism of Extreme River Discharges in the Upper Danube Basin Network," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 69, no. 4, pp. 741–764, August 2020.