Detecting Confounders in Multivariate Time Series using Strength of Causation

Yuhao Liu,[†] Chen Cui,[◇] Daniel Waxman,[◇] Kurt Butler,[◇] and Petar M. Djurić[◇]
[†]Department of Applied Mathematics and Statistics
[◇]Department of Electrical and Computer Engineering
Stony Brook University, NY 11794

***Abstract*—One of the most important problems in science is understanding causation. This problem is particularly challenging when causation has to be inferred from observational data only. A further challenge of this problem is if the observed data were generated in the presence of latent confounders. In this paper, we propose a method for detecting confounders in multivariate time series using a recently introduced concept referred to as differential causal effect (DCE). The solution is based on feature-based Gaussian processes that are not only used for estimating the DCE of the observed time series but also for estimating the latent confounders. We demonstrate the performance of the proposed method with several examples. They show that the proposed approach can detect confounders and can accurately estimate causal strengths.**

I. Introduction

In many science and engineering problems, it is of fundamental importance to infer causal relationships from data — fields as diverse as medicine [7], economics [1], social sciences [10], and machine learning [22] have an interest in causal inference. While some notions of causality are indeed statistical, such as Granger causality [5], much of modern causal inference relies on interventional and counterfactual notions [21]. These notions carry strictly more information than observational data and therefore require randomized experiments. However, these experiments are often too time-consuming, expensive, or unethical to conduct, necessitating the use of observational or quasi-experimental data.

Unfortunately, inferring causal relationships from observational data is generally ill-posed, meaning further assumptions are necessary [22, pp. 135]. One particularly common assumption, which often fails to hold in practice, is causal sufficiency; this assumption states that any variable which directly affects *at least* two other variables is observed [25, pp. 22]. If causal sufficiency is assumed but does not hold, i.e., there exists latent confounders, incorrect causal conclusions are often made. To add further to the challenges of inference, causal sufficiency is difficult to test for, and it generally requires domain knowledge to establish [25, p.123].

A wide variety of methods have been developed to address the possibility of confounders, either through their detection or by learning causal models which indicate possible confoundedness. While this current work aims at addressing the former problem, we note that much work has been conducted on the latter one. These include constraint-based methods such as fast causal inference [25, pg. 144], score-based methods [4], hybrid methods [20], or asymmetry-based methods [8] — see [26] for a recent survey of such methods. A number of these methods have been adapted or extended to the special case of time series such as ANLTSM [3]and VAR-LiNGAM [9].

The problem of detecting confounders is comparatively much less explored. One line of work into confounder detection involves deriving estimators of the "structural strength of confounding" γ, where $\gamma = 0$ corresponds to the unconfounded case and $\gamma = 1$ corresponds to the entirely confounded case. For linear Gaussian-additive noise models (LinGAMs) with a scalar confounder, the authors in [13] develop an estimate of γ using spectral techniques in high dimensions. Detection using the first moment of such a spectral measure showed superior performance in [17]. The case of LinGAMs with multivariate confounders was addressed in [14] using techniques from independent component analysis; a correction term to make the estimator consistent is provided in [24].

Another approach to the detection of confounders lies on the postulate of the algorithmic Markov condition, introduced in [12]. Under this interpretation of causality, the true causal factorization is the one which minimizes the Kolmogorov complexity of the factorized joint distribution. In [15], the minimum description length (MDL) is used as an approximation of Kolmogorov complexity, comparing the MDL of an unconfounded model to the MDL of a latent variable model (LVM) to detect confoundedness. To our knowledge, no confounder detection methods have been developed explicitly for time series.

Finally, if one is interested in a specific causal effect, some methods have employed LVMs to estimate the average causal/treatment effect, with the idea that proxies of confounders can be estimated from observed variables. For example, [19] uses variational autoencoders to estimate the average effect of a binary treatment. Meanwhile, using a slightly different set of assumptions more common to the potential outcomes framework of causality, [27] creates a framework for using factor models in estimating average causal effects.

In this paper, we propose a novel method for confounder detection in time series with additive Gaussian noise. To achieve this, we extend an existing notion of causal strength [2] to time series, and estimate the strength of any potential confounders using random feature-based Gaussian processes (GPs). We organize the rest of the paper as follows: in

Section II, we give background for causal models, GPs, and causal effect estimation. In Sections III and IV, we outline our proposed model and solution. Results for a variety of numerical experiments on simulated data are presented in Section V, before concluding in Section VI.

II. Background

A. Structure Causal Model and Latent Confounder

Consider a set of observed data $\{y_1, \ldots, y_N\}$, with an underlying cause and effect relationship. We can represent the relationship between each variable by a set of functions, which is called a structural causal model (SCM). Mathematically, we have,

$$y_i = f_i(Pa(y_i)) + \epsilon_i, \tag{1}$$

where $i = 1, 2, \cdots, N$, $Pa(\cdot)$ is the parent set of a given node and ϵ_i is independent noise or error of the model. We use the notion of parent set because we can represent the causal structure with a directed acyclic graph (DAG), with edges pointing from parents to children. The variable index is the same as the node index. By evaluating the functions f, we can get the causal structure and represent it via a DAG, with the cause-effect being edges pointing from the cause to the effect.

If a latent confounder z exists and causes a difference in the causal structure, then with the confounder z, the observed data y_i are not only a function of its parent set $Pa(y_i)$ of variables but also a function of the confounder, that is, we have $y_i = f_i(Pa(y_i), z) + \epsilon_i$.

B. Causal Strength

There are several possible ways to quantify the strength of a causal interaction [11], but in this work we will take an approach based on differential calculus. Since the derivative of a function can measure the sensitivity of the function's output to changes in the input, a natural measure of causal strength is to consider differentiation of functions in SCMs. Let $\mathbf{y}_t$ be an N dimensional multivariate time series. For simplicity, we focus on one of the observed time series, which we will simply symbolize by y_t and will denote its observed and unobserved parents by $\mathbf{x}_{t-1}$ and $\mathbf{z}_{t-1}$, respectively. We note that all the parents take their values before y_t takes its own value, which is indicated by the indices of $\mathbf{x}$ and $\mathbf{z}$. In summary, $Pa(y_t) = \{\mathbf{y}_{t-1}, \mathbf{x}_{t-1}, \mathbf{z}_{t-1}\}$, where the vector $\mathbf{y}_{t-1}$ contains all the parents of y_t that represent some of the past values of y_t, and $\mathbf{x}_{t-1}$ and $\mathbf{z}_{t-1}$ are parents that are past values of other observed and unobserved time series, respectively. If for y_t we write $y_t = f(\mathbf{y}_{t-1}, \mathbf{x}_{t-1}, \mathbf{z}_{t-1})$, we define the differential causal effect (DCE) of a single parent, e.g., of $x_{t-l,i}$ on y_t to be the partial derivative of the function with respect to $x_{t-l,i}$ [2],

$$\text{DCE}_{x_{t-l,i} \to y_t} \triangleq \frac{\partial f}{\partial x_{i,l}}, \tag{2}$$

where $x_{t-l,i}$ corresponds to the ith time series of the remaining $N-1$ time series and l is the lag of that time series, with i and l fully defining the parent. We refer to this notion of causal strength as the *direct* DCE, since it assumes that $x_{t-l,i}$ directly causes changes to y_t. More generally, $x_{t-l,i}$ might not effect y_t directly, but through a chain of causal mechanisms it exerts an influence on y_t. In this case, it is more appropriate to use the chain rule to decompose the total effect as the product of the effects along the chain. Hence, the *total* DCE $x_{t-l,i}$ on y_t is defined to be the causal effect yielded by the composition of multiple mechanisms, i.e.

$$\text{(Total) DCE}_{x_{t-l,i} \to y_t} = \frac{\partial f}{\partial Pa(y_t)} \frac{\partial Pa(y_t)}{\partial x_{t-l,i}}. \tag{3}$$

To compute the total DCE for any given interaction, we repeatedly apply (3) to derive the correct expression.

When a latent process z_t exerts an influence on another process y_t, the DCE $\partial y_t / \partial z_{t-l,i}$ will be nonzero for some lag l. Otherwise, the function f is effectively constant with respect to changes in $z_{t-l,i}$, and we cannot say that y_t depends on $z_{t-l,i}$ meaningfully. Since a reconstruction of a latent process is not unique, the magnitude of the causal strength of z_{t-i} on y_t is not generally meaningful. However, since zeroness of the causal strength does not depend on the choice of coordinates, i.e., if $\tilde{z}_{t-i}$ and z_{t-i} are two equivalent latent states, and the causal strength of $\tilde{z}_{t-i}$ is zero, then the chain rule states that

$$\frac{\partial y_t}{\partial z_{t-i}} = \frac{\partial y_t}{\partial \tilde{z}_{t-i}} \frac{\partial \tilde{z}_{t-i}}{\partial z_{t-i}} = 0 \times \frac{\partial \tilde{z}_{t-i}}{\partial z_{t-1}} = 0.$$

Thus, in principle we can use the causal strength to decide that z_{t-i} does not effect y_t. In the multivariate setting, we assert that $\mathbf{z}_{t-i}$ does not effect $\mathbf{y}_t$ when all partial derivatives are zero.

C. Gaussian processes

Gaussian processes (GPs) are a class of stochastic processes that are used in machine learning for modeling functions [23]. More specifically, let $(\mathbf{x}_t, y_t)$, $t = 1, 2, \ldots, T$, be T input-output values, where $\mathbf{y} = [y_1 \; y_2 \ldots y_T]^\top$, and $\mathbf{y} = \mathbf{f}(\mathbf{X})$, with $\mathbf{f} \in \mathbb{R}^{T \times 1}$ and $\mathbf{X} \in \mathbb{R}^{T \times d_x}$ being a matrix whose rows represent the inputs to the function $\mathbf{f}$, that is, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]^\top$, $\mathbf{y} = \mathbf{f}(\mathbf{X}) = [f(\mathbf{x}_1^\top), \ldots, f(\mathbf{x}_T^\top)]^\top$. The idea behind GPs is that function samples are jointly drawn from a Gaussian distribution. Mathematically, we have $\mathbf{f} \sim \mathcal{GP}(\mathbf{m}(\mathbf{X}), \mathbf{K}_{\boldsymbol{\theta}}(\mathbf{X}))$, where $\mathbf{m}(\mathbf{X})$ is the mean function, $\mathbf{K}_{\boldsymbol{\theta}}(\mathbf{X})$ is the covariance (kernel) function of the process, and $\boldsymbol{\theta}$, is a vector of hyperparameters of the GP.

1) Random feature-based GPs: The biggest drawback of GPs is their poor scaling, that is, GPs do not scale up well computationally with the number of input-output pairs, T. We can address this problem by resorting to an approximation by way of exploiting the concept of sparsity. One approach is based on constructing GPs with features that come from a feature space [16]. A GP with a shift-invariant kernel can be approximated using a feature space where matrix decompositions will not be required. The vector of basis functions of the feature space is comprised of trigonometric functions that are defined by

$$\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}) = \frac{1}{\sqrt{J}} [\sin \mathbf{x}^\top \mathbf{v}_1 \; \cos \mathbf{x}^\top \mathbf{v}_1 \; \cdots \; \sin \mathbf{x}^\top \mathbf{v}_J \; \cos \mathbf{x}^\top \mathbf{v}_J]^\top,$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_J]$ are random features sampled from the power spectral density of the kernel of the GP. Then the kernel function $k(\mathbf{x}, \mathbf{x}')$ can be approximated by $\phi_{\mathbf{v}}^{\top}(\mathbf{x})\phi_{\mathbf{v}}(\mathbf{x}')$ if the kernel is shift-invariant. The GP approximation is then

$$f(\mathbf{x}) \approx \phi_{\mathbf{v}}^{\top}(\mathbf{x})\boldsymbol{\theta} = [\cos(\mathbf{x}^{\top}\mathbf{V}), \sin(\mathbf{x}^{\top}\mathbf{V})]\boldsymbol{\theta}/\sqrt{J}, \quad (4)$$

where $\boldsymbol{\theta} \in \mathbb{R}^{2J\times 1}$ is a vector of parameters of the approximating model. The derivatives of the random feature-based function with respect to $\mathbf{x}$ is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \boldsymbol{\theta}^{\top}[\mathrm{diag}\left(-\sin \mathbf{V}^{\top}\mathbf{x}\right), \mathrm{diag}\left(\cos \mathbf{V}^{\top}\mathbf{x}\right)]\mathbf{V}^{\top}/\sqrt{J}. \quad (5)$$

III. Problem Formulation

Let $\mathbf{x}_t \in \mathbb{R}^{N\times 1}$ represent a vector of signals collected from a directed graph $\mathcal{G}$ from all its nodes at time t, where $x_{t,i}$ denotes the graph signal of node i at time t. The directed graph $\mathcal{G}$'s structure represents the causal relationship between each variable corresponding with the nodes in the graph. Further, we assume that the signal $x_{t,i}$ is a function of the previous data generated from all its parents. Since we investigate the injection effect from all other nodes to one specific node, then node by node, we represent the target variable $x_{t,i}$ at time t as y_t and keep its other parents $Pa(x_{t,\setminus i})$ as $\mathbf{x}$ for clarification purposes. If there is an underlying not observed process, i.e., confounding process, we represent it by $\mathbf{z}_t$. Specifically, consider the data model:

$$\mathbf{z}_t = f(\mathbf{z}_{t-l_{zz}:t-1}, \mathbf{x}_{t-l_{zx}:t-1}, y_{t-l_{zy}:t-1}) + \mathbf{u}_t, \quad (6)$$
$$\mathbf{x}_t = h(\mathbf{z}_{t-l_{xz}:t-1}, \mathbf{x}_{t-l_{xx}:t-1}, y_{t-l_{xy}:t-1}) + \mathbf{v}_t, \quad (7)$$
$$y_t = g(\mathbf{z}_{t-l_{yz}:t-1}, \mathbf{x}_{t-l_{yx}:t-1}, y_{t-l_{yy}:t-1}) + e_t, \quad (8)$$

where l_{zz}, l_{zx}, etc. are the maximum lags of past samples effecting values of the caused variables, and $\mathbf{u}_t$, $\mathbf{v}_t$, and e_t are errors modeled as zero-mean Gaussians, and $w_{i:j}$ for $w = x, y, z$ denotes $w_i, w_{i+1}, \ldots, w_j$.

The functions $f(\cdot, \ldots, \cdot)$, $h(\cdot, \ldots, \cdot)$, and $g(\cdot, \ldots, \cdot)$ are *unknown* and we assume the functions are drawn from Gaussian processes. The objective is to determine the causal strengths of given nodes to a node of interest. As a metric for causal strength we use DCE defined by

$$\mathrm{DCE}_{x_{t-l,i}\to y_t} \triangleq \frac{\partial g(\cdot,\cdot,\cdot)}{\partial x_{t-l,i}}, \quad (9)$$

where $x_{t-l,i}$ represents the causing time series and its lag.

IV. Proposed Solution

We investigate the nodes one by one, and without loss of generality, we focus on the scalar target node y_t in the remaining part of the paper. We write (6) and (8) using the form of random features as

$$\mathbf{z}_t = \mathbf{H}^{\top}\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{z}_{t-l_{zz}:t-1}, \mathbf{x}_{t-l_{zx}:t-1}, y_{t-l_{zy}:t-1}) + \mathbf{u}_t, \quad (10)$$
$$y_t = \boldsymbol{\theta}^{\top}\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{z}_{t-l_{yz}:t-1}, \mathbf{x}_{t-l_{yx}:t-1}, y_{t-l_{yy}:t-1}) + e_t, \quad (11)$$

where $\boldsymbol{\phi}_{\mathbf{v}}$ represents random vectors with $\mathbf{V} = \{\mathbf{V}_x, \mathbf{V}_y\}$, $\mathbf{H} = [\boldsymbol{\eta}^{[1]}, \boldsymbol{\eta}^{[2]}, \ldots, \boldsymbol{\eta}^{[d_x]}]$, and $\boldsymbol{\theta}$ are parameter variables. We assume that the parameter variables are all independent, i.e., the columns of $\mathbf{H}$ are independent of the other columns. The independence assumption about the parameter variables implies that the dimensions of $\mathbf{z}_t$ are conditionally independent. To do the sequential inference on the distribution of $\mathbf{H}$, $\boldsymbol{\theta}$, and $\mathbf{z}_t$, we assign prior distributions $p(\mathbf{H}), p(\boldsymbol{\theta})$, and $p(\mathbf{z}_0)$ to them and adopt the Bayesian paradigm [18].

The method for finding causal strengths in the possible presence of confounders is based on running particle filtering and Bayesian update. There are two groups of Kalman filters, and they track $\mathbf{H}_t$, and $\boldsymbol{\theta}_t$, respectively. Both Kalman filters use the estimated values of the confounder $\widehat{\mathbf{z}}_t$. The particle filter for tracking $\mathbf{z}_t$, on the other hand, uses the estimated matrices $\widehat{\mathbf{H}}_t$ and $\widehat{\boldsymbol{\theta}}_t$ to estimate the confounder time series. The Kalman filter that estimates $\boldsymbol{\theta}_t$ will produce the mean of the estimate, $\bar{\boldsymbol{\theta}}_t$, and its covariance matrix, $\boldsymbol{\Sigma}_t$, which are then used to determine the mean and variance of the desired partial derivative of y_t in terms of eq. (5). The detailed procedures and related codes can be found in [18].

V. Numerical Results

In the experiments, we considered two different cases. The first case has constant causal strength, while the second one has time-varying causal strength of the confounder. To validate that our model can detect the absence of a confounder, we added an unrelated dimension to the confounder in Synthetic Case 2 below. Both cases have only one time unit lag for the sake of easy understanding.

A. Synthetic Case 1

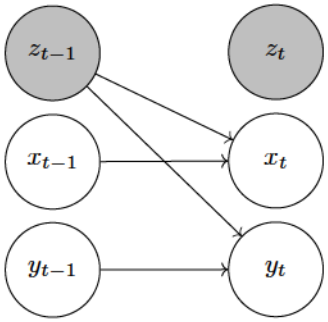

Fig. 1. Diagram for Case 1. The symbol $z_{t\ \ 1}$ is the hidden confounder while x_t and y_t are observed.

The latent state z_t contributes to both x_t and y_t, while x_t and y_t are not connected. We generated 8,000 samples by

$$z_t = \frac{1}{1+\exp\{15\sin(t/20)\}}, \quad (12)$$
$$x_t = -0.8z_{t-1} + 0.5x_{t-1} + e_t, \quad (13)$$
$$y_t = 0.5z_{t-1} - 0.8y_{t-1} + v_t, \quad (14)$$

where $e_t \sim N(0,1)$, and $v_t \sim N(0, 0.01)$. The first T_0 =1,000 samples are pre-trained, and the remaining 7,000 samples are used for real-time learning. Figure 2 presents the DCEs of y_t to the previous observations and confounder, which are

$$\frac{\partial y_t}{\partial y_{t-1}}, \frac{\partial y_t}{\partial x_{t-1}}, \text{and} \frac{\partial y_t}{\partial z_{t-1}}.$$

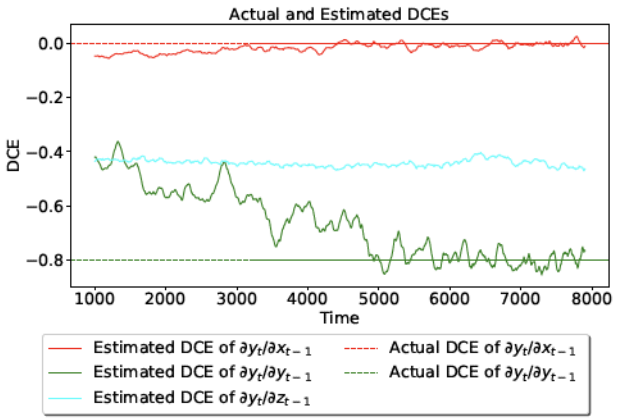

Fig. 2. Case 1: The actual and estimated DCEs of the parent nodes $y_{t\ 1}$, $x_{t\ 1}$, and $z_{t\ 1}$ to y_t.

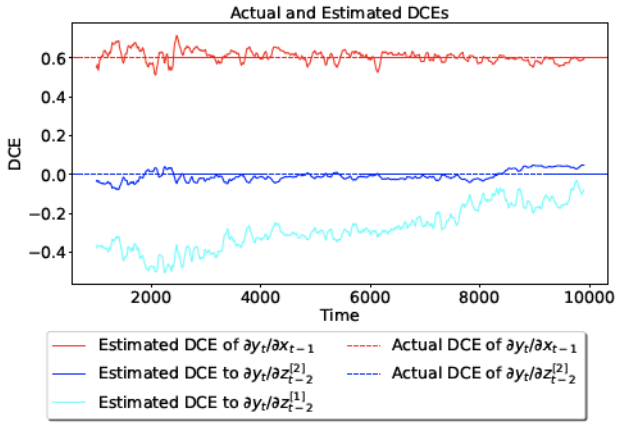

Fig. 3. Case 2: The actual and estimated DCEs of the parent nodes $\mathbf{z}_{t\ 1}$ and $x_{t\ 1}$ to y_t.

To make the lines smoother and more stable, we use moving averages of the estimated DCEs, i.e., the means of estimated DCEs among the rolling window with a fixed width. In this paper, we set the width as 100 time units. From Fig. 2, the estimated derivatives $\partial y_t/\partial y_{t-1}$ and $\partial y_t/\partial x_{t-1}$ are both around the actual derivatives. Note that the x_t and y_t have no connection. Our results show that the estimated DCE $\partial y_t/\partial x_{t-1}$ is around zero, which implies that there is no causal strength from the observed processes x_t to y_t. The estimated latent states are not unique due to the unknown f and g, but are identifiable up to scales, rotation, and mirroring on account of the properties of random feature-based GPs [6]. Consequently, we might not detect the real value of casual strengths. However, from the cyan lines in Fig. 2, the estimated DCEs of y_t to the confounder z_{t-1} are around -0.5 while the actual DCE $\partial y_t/\partial z_{t-1}$ is 0.5. Although we cannot determine whether the causation is positive or negative, the results suggest that we are not far from the absolute value of the DCEs.

B. Synthetic Case 2

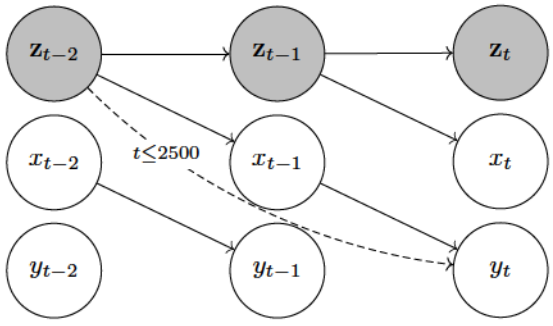

Fig. 4. Diagram for case 2. $\mathbf{z}_t$ is a two dimensional hidden process while x_t and y_t are observed.

In this experiment, we studied a case where one of the causal strengths is time-varying. The latent states $\mathbf{z}_t$ effects x_t, and both $\mathbf{z}_t$ and x_t effect y_t, where $\mathbf{x}_t$ and y_t are known while $\mathbf{z}_t$ is unknown. We generated 10,000 samples by

$$z_t^{[1]} = 0.9z_{t-1}^{[1]} + 0.5\sin(z_{t-1}^{[1]}) + u_t^{[1]}, \quad (15)$$

$$z_t^{[2]} = 0.5\sin(z_{t-1}^{[1]}) + 0.9z_{t-1}^{[2]} + u_t^{[2]}, \quad (16)$$

$$x_t = 1.2z_{t-1}^{[1]} - 0.8z_{t-1}^{[2]} + 0.8x_{t-1} + e_t, \quad (17)$$

$$y_t = 0.4z_{t-2}^{[1]} + 0.6x_{t-1} + v_t, \quad when\ t \leq 2500, \quad (18)$$

$$y_t = 0.6x_{t-1} + v_t, \quad when\ t > 2500, \quad (19)$$

where $u_t^{[1]}, u_t^{[2]} \sim N(0, 10^{-2})$, and e_t, v_t are both distributed according to $N(0, 10^{-6})$. The first T_0 =1,000 samples were pre-trained, and the remaining 9,000 samples were learned online. Figure 3 shows the DCEs of y_t to x_{t-1} and $\mathbf{z}_{t-2}$. From the figure, the estimated DCEs of $\partial y_t/\partial x_{t-1}$ and $\partial y_t/\partial z_{t-2}^{[2]}$ are around the actual values of the DCEs and are equal to 0.6 and 0, respectively. It is noteworthy that $z_{t-2}^{[2]}$ is the second dimension of the latent confounder with no effect to either x_{t-1} or y_t.

Our results show that only one of the latent dimensions effects the observations. The results suggest that our proposed model can identify the dimensions of the latent confounder. Moreover, Fig. 3 provides evidence that our method can also estimate time-varying causal strengths. The actual DCE of $\partial y_t/\partial z_{t-1}^{[1]}$ should be 0.4 before $t \leq 2500$, while it falls to zero due to the sudden disappearance of causation, as shown by (19). The cyan line in Fig. 3, representing the estimated DCEs of $\partial y_t/\partial z_{t-1}^{[1]}$, is significantly non-zero before $t = 2500$ while converging to zero after the change point. The estimated DCE cannot behave like the actual DCE that drops to zero suddenly because the Bayesian structure stores past information. One can expand our proposed Bayesian model by [28] so that the learning rate or the forgetting rate can be adjusted. The estimated latent states are identifiable up to linear transformations, which causes the estimated DCEs $\partial y_t/\partial z_{t-1}^{[1]}$ are also linearly transformed. In this case, the actual DCE of $\partial y_t/\partial z_{t-1}^{[1]}$ is 0.4 before $t = 2500$, while our estimated DCEs are around -0.4. We cannot guarantee that the

sign of causation is positive or negative, but the absolute value of the DCE is estimated closely.

VI. Summary

In this paper, we address the problem of detecting latent confounders from observed multivariate time series. We apply random feature-based Gaussian processes to (a) estimate the unknown functions that describe the relationships between the time series and (b) track the latent confounders in the hypothesized system of time series. For estimating causal strengths, we use the concept of differential causal effect. We provide simulation examples that demonstrate the ability of our approach to detect confounders.

References

[1] J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

[2] K. Butler, G. Feng, and P. M. Djurić. A Differential Measure of the Strength of Causation. *IEEE Signal Processing Letters*, 29:2208–2212, 2022.

[3] T. Chu, C. Glymour, and G. Ridgeway. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(5), 2008.

[4] M. Drton and T. S. Richardson. Iterative conditional fitting for gaussian ancestral graph models. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 130–137, 2004.

[5] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[6] G. Gundersen, M. Zhang, and B. Engelhardt. Latent variable modeling with random features. In *International Conference on Artificial Intelligence and Statistics*, pages 1333–1341. PMLR, 2021.

[7] M. A. Hernán and J. M. Robins. *Causal inference*. CRC Boca Raton, FL, 2010.

[8] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[9] A. Hyvärinen, S. Shimizu, and P. O. Hoyer. Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *Proceedings of the 25th International Conference on Machine Learning*, pages 424–431, 2008.

[10] G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[11] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324 – 2358, 2013.

[12] D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

[13] D. Janzing and B. Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.

[14] D. Janzing and B. Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *International Conference on Machine Learning*, pages 2245–2253. PMLR, 2018.

[15] D. Kaltenpoth and J. Vreeken. We are not your real parents: Telling causal from confounded using mdl. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2019.

[16] M. Lázaro-Gredilla, J. Quinonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881, 2010.

[17] F. Liu and L. Chan. Confounder detection in high-dimensional linear models using first moments of spectral measures. *Neural Computation*, 30(8):2284–2318, 2018.

[18] Y. Liu, M. Ajirak, and P. Djuric. Sequential estimation of gaussian process-based deep state-space models. *arXiv preprint arXiv:2301.12528*, 2023.

[19] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 30, 2017.

[20] J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379. PMLR, 2016.

[21] J. Pearl. *Causality*. Cambridge University Press, 2009.

[22] J. Peters, D. Janzing, and B. Scholkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. MIT Press, 2017.

[23] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2005.

[24] L. Rendsburg, L. C. Vankadara, D. Ghoshdastidar, and U. von Luxburg. A consistent estimator for confounding strength. *arXiv preprint arXiv:2211.01903*, 2022.

[25] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

[26] M. J. Vowels, N. C. Camgoz, and R. Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

[27] Y. Wang and D. M. Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[28] P.-S. Wu and R. Martin. A comparison of learning rate selection methods in generalized Bayesian inference. *Bayesian Analysis*, 18(1):105–132, 2023.