Nonlinear Causal Discovery for High-Dimensional Deterministic Data

Yan Zeng, Zhifeng Hao, *Member, IEEE*, Ruichu Cai, *Member, IEEE*, Feng Xie, Libo Huang, *Student Member, IEEE*, and Shohei Shimizu

***Abstract*— Nonlinear causal discovery with high-dimensional data where each variable is multidimensional plays a significant role in many scientific disciplines, such as social network analysis. Previous work majorly focuses on exploiting asymmetry in the causal and anticausal directions between two high-dimensional variables (a cause–effect pair). Although there exist some works that concentrate on the causal order identification between multiple variables, i.e., more than two high-dimensional variables, they do not validate the consistency of methods through theoretical analysis on multiple-variable data. In particular, based on the asymmetry for the cause–effect pair, if model assumptions for any pair of the data are violated, the asymmetry condition will not hold, resulting in the deduction of incorrect order identification. Thus, in this article, we propose a causal functional model, namely high-dimensional deterministic model (HDDM), to identify the causal orderings among multiple high-dimensional variables. We derive two candidates' selection rules to alleviate the inconvenient effects resulted from the violated-assumption pairs. The corresponding theoretical justification is provided as well. With these theoretical results, we develop a method to infer causal orderings for nonlinear multiple-variable data. Simulations on synthetic data and real-world data are conducted to verify the efficacy of our proposed method. Since we focus on deterministic relations in our method, we also verify the robustness of the noises in simulations.**

***Index Terms*— Causal ordering, deterministic relations, high-dimensional data, nonlinear causal discovery.**

Manuscript received 24 November 2020; revised 9 June 2021; accepted 11 August 2021. Date of publication 3 September 2021; date of current version 3 May 2023. This work was supported in part by the NSFC-Guangdong Joint Fund under Grant U1501254; in part by the Natural Science Foundation of China under Grant 61876043 and Grant 61472089; in part by the Natural Science Foundation of Guangdong under Grant 2014A030306004 and Grant 2014A030308008; in part by the Science and Technology Planning Project of Guangdong under Grant 201902010058. The work of Yan Zeng was supported by China Scholarship Council (CSC). The work of Shohei Shimizu was supported in part by ONR under Grant N00014-20-1-2501 and in part by KAKENHI under Grant 20K11708. *(Corresponding authors: Zhifeng Hao; Ruichu Cai.)*

Yan Zeng is with the School of Computer, Guangdong University of Technology, Guangzhou 510006, China, and also with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: yanazeng013@gmail.com).

Zhifeng Hao is with the College of Science, Shantou University, Shantou, Guangdong 515063, China (e-mail: zfhao@gdut.edu.cn).

Ruichu Cai is with the School of Computer, Guangdong University of Technology, Guangzhou 510006, China, and also with Pazhou Lab, Guangzhou 510006, China (e-mail: cairuichu@gdut.edu.cn).

Feng Xie is with the School of Mathematical Sciences, Peking University, Beijing 100084, China (e-mail: xiefeng009@gmail.com).

Libo Huang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100084, China (e-mail: www.huanglibo@gmail.com).

Shohei Shimizu is with the Faculty of Data Science, Shiga University, Hikone 522-8522, Japan, and also with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan (e-mail: shohei-shimizu@biwako.shiga-u.ac.jp).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3106111.

Digital Object Identifier 10.1109/TNNLS.2021.3106111

I. Introduction

CAUSAL discovery on observational data has attracted extensive attention from researchers on different domains in the past decades [1]–[4]. Although traditional methods, such as PC [5], IC [6], and GES [7], have been used in a range of fields, they cannot distinguish causally distinct models that contain Markov equivalence classes, i.e., a set of causal structures satisfying the same conditional independence. Recently, functional causal models (FCMs) have been proposed, which can distinguish the same equivalence classes and uniquely identify the causal structure, such as LiGNAM [8], ANM [9], and PNL [10]. However, few of these methods can cope with the deterministic relationships between nonlinear high-dimensional variables.[1]

The fact that deterministic relations exist in real-world settings demonstrates that priors, as are usually assumed for the causal networks, are problematic, and constructing good priors is difficult [11]. Thus, instead of defining the priors explicitly, here, we focus on analyzing the deterministic relationships. Furthermore, causal discovery for high-dimensional random variables is a quite practical issue in various fields, e.g., climate research, and financial market analysis [12]. For instance, if we infer the causal relations between geographic position ($\boldsymbol{X}$) and precipitation ($\boldsymbol{Y}$), then they may be defined as $\boldsymbol{X} = \{\text{altitude}, \text{longitude}, \text{latitude}\}$ and $\boldsymbol{Y} = \{\text{prec. in Jan}, \ldots, \text{prec. in Dec}\}$. If we intend to recover the causal relations between the daily stock returns in different regions of the world, i.e., America ($\boldsymbol{X_1}$), Europe ($\boldsymbol{X_2}$), and Asia ($\boldsymbol{X_3}$), then we may define $\boldsymbol{X_1} = \{\text{BSESN}, \text{DJI}, \text{GSPC}, \text{IXIC}\}$, $\boldsymbol{X_2} = \{\text{BUK100P}, \text{FCHI}\}$, and $\boldsymbol{X_3} = \{\text{HSI}, \text{N225}\}$, where the abbreviations in $\boldsymbol{X_1}$ stand for the stock returns in USA, those in $\boldsymbol{X_2}$ stand for U.K. and France, and those in $\boldsymbol{X_3}$ stand for China and Japan, respectively. Note that the aforementioned variables, e.g., $\boldsymbol{X_1}$ or $\boldsymbol{Y}$, are all multidimensional. However, inferring causal relations for such data is challenging, especially when the high-dimensional variables are deterministic.

[1] Here, "high-dimensional variables" are those variables that are multidimensional. A high-dimensional variable is also usually called a variable group in causal discovery.

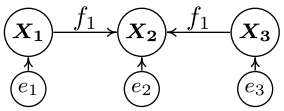

Fig. 1. Simple case where incorrect identification may happen when some pairs in the data do not follow the assumptions made for the two-variable methods. Suppose that we have three high-dimensional variables $\boldsymbol{X_1}, \boldsymbol{X_2}$, and $\boldsymbol{X_3}$. f_1 represents nonlinear relationships between variables. If we assume the cause–effect pair that follows the deterministic model with no noises such as the KTR method [16], we have $\boldsymbol{X_1} = \boldsymbol{X_1}$, $\boldsymbol{X_3} = \boldsymbol{X_3}$, and $\boldsymbol{X_2} = f_1(\boldsymbol{X_1}, \boldsymbol{X_3})$. It can be seen that $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, abbreviated as $\{\boldsymbol{X_1}, \boldsymbol{X_3}\}$, do not follow deterministic relations anymore, which means that they do not own the asymmetry. Same conclusions can be drawn for the pairs, including $\{\boldsymbol{X_3}, \boldsymbol{X_2}\}$ and $\{\boldsymbol{X_1}, \boldsymbol{X_3}\}$. Hence, using the cause–effect asymmetry toward this model to identify orderings will induce errors. Similarly, if we assume that the cause–effect pair follows the ANM model with additive noises [9], we have, $\boldsymbol{X_1} = e_1$, $\boldsymbol{X_3} = e_3$ and $\boldsymbol{X_2} = f_1(\boldsymbol{X_1}, \boldsymbol{X_3}) + e_2$, where e_i is the noise of $\boldsymbol{X_i}$. It implies that both $\{\boldsymbol{X_1}, \boldsymbol{X_2}\}$ and $\{\boldsymbol{X_1}, \boldsymbol{X_3}\}$ do not follow the ANM model; consequently, some errors would also arise in this case.

Several authors have sought solutions from a perspective of "asymmetry" on a cause–effect pair first [13]–[15] to tackle the problem of causal discovery with high-dimensional variables. This "asymmetry" is given rise to by the "independent mechanisms of nature" that the distribution of the cause and the conditional distribution of the effect given the cause are generated independently, which can be exploited to identify the causal direction. Motivated by this asymmetry, the linear trace (LTR) method [11] has become one of the most enlightening approaches for the linearly coupled high-dimensional cause–effect pairs. Based on LTR, Zscheischler *et al.* [12] extended it to the noisy case on the condition of a sparsity constraint, and Chen *et al.* [16] relaxed the linearity assumption, settling the nonlinear causality problem with the kernel-based method, called kernelized trace (KTR) method. Unfortunately, previous methods aim at inferring causal directions between two high-dimensional variables, but not between multiple high-dimensional variables. This is a serious limitation since such multiple-variable data ubiquitously exist in real-world scenarios [12], [17].

For the case of more than two high-dimensional variables, there still exists some work [18]–[20]. Entner and Hoyer [18] focused on uncovering connections in linear models by exploiting non-Gaussianity. However, the fact of linearity assumption of the causal mechanism limits its applicability. For nonlinear models, Chen *et al.* [19] proposed a generalized method, namely Hilbert space embedding-based method (EMD), and Liu *et al.* [20] took the use of free probability, unveiling the causal order in multiple variables. However, both methods pay much attention to theoretically exploiting the "asymmetry" in a cause–effect pair and directly extend it into the multiple-variable data. In other words, it will not be straightforward to apply methods for two-variable data on multiple-variable data. It is because if two high-dimensional variables are taken from multiple-variable data, the pair may not follow the assumptions made for the two-variable methods. They are lack in theoretical guarantees for the multiple-variable methods. Thus, incorrect identification of directions will be arisen. A simple case is provided in Fig. 1 where incorrect identification may happen in the multiple-variable data.

Thus, in this article, we propose an integrated method to avoid this phenomenon. In particular, we first define a causal functional model, namely the multiple high-dimensional deterministic model (HDDM). Interestingly, based on this model, we obtain an essential characteristic, with which we derive candidates' selection rules. We achieve that more pairs in the multiple high-dimensional variables will own the asymmetry so that the order identification accuracy is improved. Moreover, for the nonlinear problem, we follow the line of [11] and [16], using the favorable property of the reproducing kernel Hilbert space (RKHS). This property is to map the data into RKHS such that the nonlinear relations become linear ones as close as possible [21]. The main contributions of this article are summarized as follows.

1) We present HDDM, which owns an available characteristic and facilitate identifying the causal orderings for multiple high-dimensional variables.
2) We derive two candidates' selection rules and propose an integrated method to identify the causal orderings. Furthermore, we give theoretical results to assure the consistency of our method.
3) We do not assume any specific form for the nonlinear functions, provided that the parents' effects are additive. Though we assume that the relation is deterministic, we also discuss that our method is also competent to deal with noisy cases.

This article is structured as follows. Section II reviews the related work, while Section III describes our basic model and problem formalization. In Section IV, we propose an integrated method to estimate the model and present its theoretical analysis. Section V shows the experimental results on synthetic data and real-world data. Finally, we present the conclusions in Section VI.

II. Related Work

In this section, before reviewing the existing state-of-the-art methods for high-dimensional causal discovery of deterministic relationships, we would introduce some basic methods for bivariate causal discovery relying on the independence of cause and mechanism postulate [4], [22]. Information geometric causal inference (IGCI) is used to infer the causal direction between 1-D cause and effect variables, which satisfy deterministic relationships [14], [15]. IGCI is based on the postulate that the distribution from cause to effect and the distribution of the cause are chosen independently, while in the anticausal direction, the distribution from effect to cause and the distribution of the effect are dependent in a certain sense. This postulate characterizes the asymmetry, which aids IGCI in determining the causal direction [4], [23]. Employing the same independence postulate, Tagasovska *et al.* [24] made use of the quantile regression to distinguish the cause from the effect. Blöbaum *et al.* [25] compared the mean-regression errors and develops the regression error-based causal inference (RECI) method to infer the causal relation between two variables. Shajarisales *et al.* [26] also took the best of this postulate and proposed an approach to infer a cause from its effect in deterministic linear dynamical systems. Various extensions include [27], [28], [29], and [30]. Though the above-mentioned methods achieve satisfactory performance between nonlinear

coupled 1-D cause–effect pairs, it is not applicable in high-dimensional data.

For the high-dimensional variables, we first consider the LTR method [11]. It analyzes the linear deterministic model, $\mathbf{Y} = A\mathbf{X}$, inferring the causal direction between high-dimensional cause–effect pairs, $\mathbf{X}$ and $\mathbf{Y}$. The idea is to harness an asymmetry between the distributions of cause and effect that occurs when the covariance matrix of the cause $\Sigma_{\mathbf{X}}$ and the structure matrix A mapping the cause to the effect correspond to independent mechanisms of nature. Based on this idea, LTR thus fulfills the trace condition [13], which consequently renders it possible to identify the causal direction. Similarly, Janzing and Schölkopf [31] considered linear models where the predictor variable is multidimensional and it influences one target quantity. Their proposed method can detect whether these two variables are causal or not; Liu and Chan [32] utilized the first moments of spectral measures to detect the confounder in high-dimensional linear models, and Janzing [33] also considered the case where a multidimensional variable influences a target variable. These methods are inspiring but also exist some drawbacks: 1) when faced with complex nonlinear data, they may fail to cope with it, and 2) inability to cope with multiple high-dimensional data would hinder their usage in various real-world situations since multiple-variable data are omnipresent.

To settle down the problem of linearity assumption of LTR, Chen *et al.* [16] proposed a nonlinear causal discovery method, called KTR method, for two high-dimensional variables. The basic idea is to map the hypothetical cause $\mathbf{X}$ into RKHS using an empirical kernel map functions, such that in RKHS, the nonlinear relations between the cause $\mathbf{X}$ and the effect $\mathbf{Y}$ would become simpler linear ones. They justify that the trace condition also holds in nonlinear relations for the empirical kernel map so that KTR can identify the causal direction correctly and soundly. However, they do not give how to extend their work to the multiple-variable case, either, which is the same as LTR method [11]. Various extensions and modifications on high-dimensional bivariate causal discovery have also been developed (e.g., [34]–[38]).

By incorporating the kernel mean embeddings, Chen *et al.* [19] accessed and characterized an asymmetry based on the key idea that the independence between the distribution of the cause and that of effect given the cause induces an uncorrelated condition, motivated by the LTR method. Fulfilling this condition, a method called EMD is proposed based on the complexity matrix. Their method deals with nonlinear high-dimensional variables, not only inferring causal directions between cause–effect pairs but also estimating a causal ordering among multiple variables. With the identical goals, Liu and Chan [20] restored the theory of free probability for causal discovery to exploit asymmetry and improve the EMD's performance. This improved method is called FI. Rather than calculating the complexity matrix, it makes the best of freeness condition, i.e., covariance matrix of the RKHS embedding of the distribution of $\mathbf{X}$ is free independent with the covariance matrix of the RKHS embedding of the distribution of $\mathbf{Y}$ given $\mathbf{X}$ [20], [39], [40]. In principle, frameworks of EMD and FI for inferring the causal ordering among multiple high-dimensional variables are identical. Besides, Bazin *et al.* [41] introduced the formal concept analysis into causal discovery and further used an algorithmic approach for inferring causal directions for high-dimensional variables by the Kolmogorov complexity. Though these methods achieve the identification of a causal ordering among multiple high-dimensional variables, they lack theoretical analysis for such algorithms they proposed since they majorly focus on exploring the asymmetry among two cause–effect variables.

Moreover, Entner and Hoyer [18] extended current appealing methods [42], [43] and developed a collection of methods to identify the causal ordering among linear multiple high-dimensional variables. They are group pairwise measure and group trace method [GroupDirectLiNGAM (GDL)] [18]. However, they cannot be applied to nonlinear cases.

III. Problem Formalization

Following the notations in [19] and [20], we let $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ denote the observed multiple variables, where each $\boldsymbol{X_i}$ is m_i-dimensional. Here, $\boldsymbol{X}_i$ is also called a variable group in causal discovery. We assume that the data satisfy the following assumptions.

A1: The value of the variable $\boldsymbol{X_i}$ is a nonlinear function of the values on the earlier variables $\boldsymbol{X_j}$, i.e.,

$$\boldsymbol{X_i} = \sum_{K(j)<K(i)} \boldsymbol{f_{ij}}(\boldsymbol{X_j}) \tag{1}$$

where $\boldsymbol{f_{ij}}$ can be nonlinear functions or linear functions and $K(i)$ is the causal order of $\boldsymbol{X_i}$. Note that the earlier variables' effects are additive.

A2: The distribution of cause variables $\boldsymbol{X_j}$ is independent with the generation mechanism $\boldsymbol{f_{ij}}$.

For instance, for the causal structure $\boldsymbol{X_1} \rightarrow \boldsymbol{X_2} \leftarrow \boldsymbol{X_3}$, we assume that $\boldsymbol{X_2}$ is generated via

$$\begin{aligned}\boldsymbol{X_2} &= \boldsymbol{f_{21}}(\boldsymbol{X_1}) + \boldsymbol{f_{23}}(\boldsymbol{X_3}) \\ &= \left[f_{21}^{(1)T}(\boldsymbol{X_1}), \ldots, f_{21}^{(m_2)T}(\boldsymbol{X_1})\right]^T \\ &\quad + \left[f_{23}^{(1)T}(\boldsymbol{X_3}), \ldots, f_{23}^{(m_2)T}(\boldsymbol{X_3})\right]^T.\end{aligned}$$

In particular, when $\boldsymbol{f_{21}}$ and $\boldsymbol{f_{23}}$ are both linear functions, the generation mechanism of $\boldsymbol{X}_2$ is deduced to: $\boldsymbol{X}_2 = \boldsymbol{f_{21}}(\boldsymbol{X_1}) + \boldsymbol{f_{23}}(\boldsymbol{X_3}) = \boldsymbol{A}_{21}\boldsymbol{X_1} + \boldsymbol{A}_{23}\boldsymbol{X_3}$, where $\boldsymbol{A}_{21}$ and $\boldsymbol{A}_{23}$ are two structure matrices. Note that the assumption A2 implies that the distribution of the cause and the conditional distribution of the effect given cause are generated independently, which corresponds to two independent natural processes and contains no information about each other. It is widely used in causal discovery [4], [14], [24]. Based on the model's assumptions, it is possible to derive the asymmetry for any two variables in a causal network so that we can identify the causal orderings.

Definition 1 (HDDM): A model with Assumptions A1 and A2 is called the HDDM.

Estimating the causal ordering of this model is acknowledged to be interesting but challenging [11], [19]. One may need to address the following tough problems.

1) The relations might be deterministic and there is no noise that can provide hints for the direction identification.
2) Because of the high dimensionality of each variable, the quantities of interest may not be estimated reliably, which could lead to unreliable results. Furthermore, the nonlinear functions might be so complicated that it is hard to handle.
3) When there are more than two high-dimensional variables, it is difficult for any pairs to still satisfy the deterministic relationship.

To tackle Problem 1), based on the IGCI and KTR in [14] and [16], focusing on deterministic cases, we assume that the nonlinear functions $\boldsymbol{f_{ij}}$ are generated by a mechanism that is independent with the distribution of the cause $\boldsymbol{X_j}$. With such an assumption, we can do inference in deterministic cases as in [16]. To tackle Problem 2), there is a popular technique currently, i.e., kernel methods [44]. This is based on an intuition that mapping the data from the original space into an RKHS can establish a condition that nonlinear functions in the original space become simple linear ones in RKHS. Such intuition has been demonstrated powerful in handling nonlinearity as well as high dimensionality, including [16] and [45]. For example, Chen *et al.* [16] used this intuition to exploit an asymmetry of the joint distribution to distinguish the high-dimensional cause and effect. These were the works for two high-dimensional variables. However, when it involves more variables, we cannot directly adopt the aforementioned method since it is not credible for any two variables to follow the assumptions made for the two-variable models, e.g., deterministic relations [16] or ANM [9] assumption. Otherwise, it will bring about improper identification of causal ordering. Here, comes Problem 3) and its solutions, which lie at the heart of our proposed method. To tackle Problem 3), we employ a simple trick from the HDDM model, and since our method is based on the deterministic KTR for two-variable data [16], we subsequently find more pairs to satisfy deterministic relationships with this simple trick. We achieve how to identify an exogenous variable and how to guarantee the method is consistent, which are all described in Section IV.

IV. Theories and Methods

In this section, we first briefly review the method between a high-dimensional cause–effect pair. Next, we generalize the cause–effect pair into cases with more than two high-dimensional variables, i.e., multiple high-dimensional variables. In particular, we exhibit a simple trick of the HDDM and derive two candidates' selection rules to realize the identification of exogenous variables. In addition, the consistency of our model is guaranteed theoretically as well. Finally, we present our method.

A. Previous Method of Two High-Dimensional Variables

KTR method [16] exploits a kind of asymmetry between a cause–effect pair to identify its direction, and such asymmetry is originated from the independence between the cause and the generation mechanism. Specifically, suppose that we are given two nonlinear high-dimensional variables $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ with dimensionality of m_i and m_j, respectively. The ground truth is that $\boldsymbol{X_i}$ is generated deterministically as a nonlinear function of $\boldsymbol{X_j}$, i.e., $\boldsymbol{X_i} = \boldsymbol{f_{ij}}(\boldsymbol{X_j})$, where $\boldsymbol{f_{ij}}$ are injective and nonlinear (or linear) functions. KTR utilized the idea that constructing a mapping from the original space to RKHS can render the nonlinear relations between $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ that become linear ones, in which way it could enjoy the properties of the LTR method [11], i.e.,

$$\boldsymbol{X_i} = \boldsymbol{f_{ij}}(\boldsymbol{X_j}) = A \cdot \Psi(\boldsymbol{X_j}) \tag{2}$$

where A is a connection structure matrix which depends on the nonlinear function $\boldsymbol{f_{ij}}$ and $\Psi(\boldsymbol{X_j})$ is the image of $\boldsymbol{X_j}$ after employing an empirical kernel mapping. Due to the independence between $\boldsymbol{f_{ij}}$ and the distribution of $\boldsymbol{X_j}$, it proves that the LTR condition still holds for the causal direction while violated in the anticausal direction. In other words, in the causal direction $\boldsymbol{X_j} \to \boldsymbol{X_i}$, if the model assumption is satisfied, then the following equation will hold:

$$\tau_{m_j}\left(A\,\Sigma_{\Psi(\boldsymbol{X_j})}A^T\right) \approx \tau_{m_j}(AA^T)\tau_N\left(\Sigma_{\Psi(\boldsymbol{X_j})}\right) \tag{3}$$

where $\tau_{m_j}(\cdot) = \mathrm{tr}(\cdot)/m_j$, $\Sigma_{\Psi(\boldsymbol{X_j})}$ is the covariance matrix of $\Psi(\boldsymbol{X_j})$, and N is the sample size. Besides, in the anticausal direction, when assuming that $\boldsymbol{f_{ij}}$ is injective and mapping $\boldsymbol{X_i}$ into $\Phi(\boldsymbol{X_i})$ with the empirical kernel map, i.e., $\boldsymbol{X_j} = B \cdot \Phi(\boldsymbol{X_i})$, then the LTR condition will not hold any more

$$\tau_{m_i}\left(B\,\Sigma_{\Psi(\boldsymbol{X_i})}B^T\right) < \tau_{m_i}(BB^T)\tau_N\left(\Sigma_{\Psi(\boldsymbol{X_i})}\right). \tag{4}$$

It is this asymmetry that aids in identifying the direction. Consequently, the statistics of interest are estimated as follows:

$$\begin{aligned}\tau_{m_j}\left(\hat{A}\,\hat{\Sigma}_{\Psi(\boldsymbol{X_j})}\hat{A}^T\right) &= \frac{1}{N}\tau_{m_j}\left(\hat{A}\boldsymbol{G_j}\hat{A}^T\right) = \frac{1}{N}\tau_{m_j}\left(\boldsymbol{G_j}^2\hat{R}^T\hat{R}\right)\\ \tau_{m_j}\left(\hat{A}\hat{A}^T\right) &= \tau_{m_j}\left(\boldsymbol{G_j}\hat{R}^T\hat{R}\right)\\ \tau_N\left(\hat{\Sigma}_{\Psi(\boldsymbol{X_j})}\right) &= \frac{1}{N}\tau_N\left(\boldsymbol{G_j}\right)\end{aligned} \tag{5}$$

where $\boldsymbol{G_j}$ is the kernel Gram matrix of $\boldsymbol{X_j}$, $\hat{R} = \boldsymbol{X_i}(\boldsymbol{G_j} + \lambda \boldsymbol{I_N})^{-1}$, and λ is the regularization parameter. $\hat{B}$ and $\hat{\Sigma}_{\Phi(\boldsymbol{X_i})}$ can be estimated in a similar way.

Hence, it concludes a scale-invariant measure

$$\begin{aligned}\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} := \log\tau_{m_j}\left(\hat{A}\,\hat{\Sigma}_{\Psi(\boldsymbol{X_j})}\hat{A}^T\right) - \log\tau_{m_j}\left(\hat{A}\hat{A}^T\right)\\ - \log\tau_N\left(\hat{\Sigma}_{\Psi(\boldsymbol{X_j})}\right).\end{aligned} \tag{6}$$

It proves that (6) holds for the causal direction, i.e., $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} \approx 0$, while it is violated for the anticausal direction, i.e., $\Delta_{\Psi(\boldsymbol{X_i})\to\boldsymbol{X_j}} \leq 0$.

B. Method of Multiple High-Dimensional Variables

Based on the work of deterministic KTR [16], we generalize the cause–effect pairs into cases with more than two high-dimensional variables. The major problem we are faced with is how to guarantee that any pairs in a multivariable dataset still follow the model assumptions of KTR, i.e., deterministic relationships. As shown in Fig. 1, $\{\boldsymbol{X_3}, \boldsymbol{X_2}\}$ and $\{\boldsymbol{X_1}, \boldsymbol{X_3}\}$ pairs do not follow the KTR or ANM assumptions. In these cases, if we employ the methods which uncover the direction between

cause and effect variables into the multivariable data, incorrect identification may occur. In other words, how to alleviate the effects when the cause–effect pairs do not initially follow the deterministic relationships becomes the challenge. This is a significant problem since [11] has mentioned it is a sanity check for complex causal networks. Interestingly, our proposed HDDM model owns an essential characteristic, which enables us to handle such a significant problem.

Our basic idea is based on a simple trick of the model, substantially an available characteristic of the HDDM, which is demonstrated in Theorem 1.

Theorem 1: Assume that data $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ strictly follow the HDDM (1). Then, any two dependent variables $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ with $j \neq i$ and $K(j) < K(i)$ have deterministic relationships and obtain the asymmetry through transforming $\boldsymbol{X_j}$ to $\boldsymbol{X_{\overline{j}}}$ shown as follows:

$$\boldsymbol{X_i} = \sum_{K(j)<K(i)} \boldsymbol{f_{ij}}(\boldsymbol{X_j}) = \hat{\boldsymbol{f}}\left(\boldsymbol{X_{\overline{j}}}\right) \tag{7}$$

where $\hat{\boldsymbol{f}}$ are injective and nonlinear (or linear) functions and $\boldsymbol{X_{\overline{j}}}$ is a union variable that concatenates the parental and ancestral variables of $\boldsymbol{X_i}$, including $\boldsymbol{X_j}$.

Proof: We prove the theorem from two aspects: 1) there is only one variable satisfying $K(j) < K(i)$ and 2) there is more than one variable satisfying $K(j) < K(i)$.

First, assume that there is only one variable $\boldsymbol{X_j}$ that satisfies $K(j) < K(i)$. In this case, it is easy to see that $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ are deterministic since $\boldsymbol{X_i} = \boldsymbol{f_{ij}}(\boldsymbol{X_j}) = \hat{\boldsymbol{f}}(\boldsymbol{X_{\overline{j}}})$ where $\boldsymbol{X_{\overline{j}}} = \boldsymbol{X_j}$.

Second, assume that there is more than one variable that satisfies $K(j) < K(i)$, and specifically, assume that there are t ($t > 1$) variables $[\boldsymbol{X_1}, \ldots, \boldsymbol{X_t}], 1, \ldots, t \neq i$ satisfying the condition. Hence, we have

$$\begin{aligned} \boldsymbol{X_i} &= \sum_{j=1}^{t} \boldsymbol{f_{ij}}(\boldsymbol{X_j}) \\ &= \sum_{j=1}^{t} \boldsymbol{f_{ij}}\left(\boldsymbol{A_j}\boldsymbol{X_{\overline{j}}}\right) \\ &= \hat{\boldsymbol{f}}\left(\boldsymbol{X_{\overline{j}}}\right) \end{aligned} \tag{8}$$

where

$$\begin{aligned} &\boldsymbol{X_{\overline{j}}} = [\boldsymbol{X_1}, \ldots, \boldsymbol{X_t}]^T; \\ &\boldsymbol{A_j} = \left[\boldsymbol{0}_{m_j \times m_1}, \ldots, \boldsymbol{I}_{m_j \times m_j}, \ldots, \boldsymbol{0}_{m_j \times m_t}\right]. \end{aligned}$$

$\boldsymbol{0}_{m_j \times m_1}$ is an $m_j \times m_1$ zero matrix, while $\boldsymbol{I}_{m_j \times m_j}$ is an $m_j \times m_j$ identity matrix. From the third equality in (8), we see that $\boldsymbol{X_i}$ and $\boldsymbol{X_{\overline{j}}}$ share a deterministic relationship. To be concluded, for any pair $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ with $K(j) < K(i)$, transforming $\boldsymbol{X_j}$ into $\boldsymbol{X_{\overline{j}}}$ renders them follow the deterministic relationships. Besides, since $\hat{\boldsymbol{f}}$ are injective, according to the asymmetry of KTR method [16], we know that $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ still obtain the asymmetry, which can be utilized for ordering identification of causal networks. Thus, the theorem is proven. □

In other words, this theorem guarantees that any pairs in the multivariable data can be transformed to have deterministic relationships so that the asymmetry still holds to facilitate us in identifying the causal orderings of complex networks. Subsequently, comes the next problem: How to make the transformations toward $\boldsymbol{X_j}$ to get $\boldsymbol{X_{\overline{j}}}$ so that $\boldsymbol{X_i}$ and $\boldsymbol{X_{\overline{j}}}$ have deterministic relations? Hence, in the following, we derive two kinds of candidates' selection rules to settle this transformation problem.

Rule 1: If there is no prior information, for cause–effect pairs $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$, we select $\boldsymbol{X_{\overline{j}}} = \boldsymbol{X_j}$. If they do not follow deterministic relationships, we select $\boldsymbol{X_{\overline{j}}} = \boldsymbol{D} \backslash \boldsymbol{X_i}$, where $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ denote the observed high-dimensional variables.

Rule 2: If there is prior information, e.g., assume that variables $\boldsymbol{X_k}$ are exogenous, for cause–effect pairs $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$, we select $\boldsymbol{X_{\overline{j}}} = [\boldsymbol{X_j}, \boldsymbol{X_k}]^T$.

For Rule 1, if we have no prior information to be used in the identification of exogenous variables, we would first utilize $\boldsymbol{X_i}$ and $\boldsymbol{X_{\overline{j}}} = \boldsymbol{X_j}$. Employing the KTR method enables us to determine whether they follow deterministic relations or not. If $\boldsymbol{X_i}$ and $\boldsymbol{X_{\overline{j}}}$ follow deterministic relationships, $\Delta_{\Psi(\boldsymbol{X_{\overline{j}}}) \to \boldsymbol{X_i}} \approx 0$ in (6) will hold for the causal direction while violated for the anticausal direction, i.e., $\Delta_{\Psi(\boldsymbol{X_i}) \to \boldsymbol{X_{\overline{j}}}} \leq 0$. Meanwhile, if either $\Delta_{\Psi(\boldsymbol{X_{\overline{j}}}) \to \boldsymbol{X_i}} \approx 0$ or $\Delta_{\Psi(\boldsymbol{X_i}) \to \boldsymbol{X_{\overline{j}}}} \leq 0$ does not hold, it means that $\boldsymbol{X_i}$ and $\boldsymbol{X_{\overline{j}}}$ are not deterministic. If they do not satisfy the deterministic relationships, we would transform $\boldsymbol{X_j}$ into $\boldsymbol{X_{\overline{j}}}$ that concatenates all variables except $\boldsymbol{X_i}$. With this type of transformation, the exogenous variables can enjoy the property that the interests of statistics Δ are approximately equal to zeros for all the other variables, as shown in Theorem 2, which helps in the order identification task. For Rule 2, if we obtain prior information, e.g., part of the causal ordering K, we would choose to compute the interests between $\boldsymbol{X_i}$ and a transformed $\boldsymbol{X_{\overline{j}}}$, where $\boldsymbol{X_{\overline{j}}}$ concatenates the exogenous variables in K. The derivation of Rule 2 is to avoid the process of removing the effects of the exogenous variable on the other variables. The detailed applicability of this rule is referred to Theorem 3 and examples of how to use Rules 1 and 2 can be seen in Section IV-C.

With these two selection rules to facilitate every pair in the data in satisfying the deterministic relationships, the next problems we meet are how to identify an exogenous variable and how to guarantee the method is consistent, i.e., how to identify exogenous variables iteratively. Here, we present two theorems.

Theorem 2: Assume that data $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ strictly follow the HDDM (1). Denote by $\Delta_{\Psi(\boldsymbol{X_j}) \to \boldsymbol{X_i}}$ in (6) the measure of causal direction. Then, $\boldsymbol{X_j}$ is exogenous if and only if $\Delta_{\Psi(\boldsymbol{X_j}) \to \boldsymbol{X_i}} \approx 0$ holds for all $i \neq j$.

In other words, $\boldsymbol{X_j}$ is exogenous if and only if for all $i \neq j$, $\Delta_{\Psi(\boldsymbol{X_j}) \to \boldsymbol{X_i}}$ is always the closest to zero, compared with $\Delta_{\Psi(\boldsymbol{X_i}) \to \boldsymbol{X_j}}$.

Proof: First, assume that $\boldsymbol{X_j}$ is exogenous, and we prove that $\Delta_{\Psi(\boldsymbol{X_j}) \to \boldsymbol{X_i}} \approx 0$ holds for all $i \neq j$. Due to the derivation of Theorem 1, we conclude that if the number of the exogenous variables is 1, i.e., $\boldsymbol{X_j}$ is the only exogenous variable, any variable $\boldsymbol{X_i}, i \neq j$ in the data $\boldsymbol{D}$ is dependent with $\boldsymbol{X_j}$, so they will satisfy the deterministic relationships. Hence, in this case, according to (3) of the KTR method [16], they will satisfy $\Delta_{\Psi(\boldsymbol{X_j}) \to \boldsymbol{X_i}} \approx 0$. If the number of exogenous variables is more than 1, we prove that $\Delta_{\Psi(\boldsymbol{X_j}) \to \boldsymbol{X_i}} \approx 0$ still

holds for both other exogenous and nonexogenous variables. First, we know that other exogenous variables are independent with $\boldsymbol{X_j}$. Therefore, the connection matrix $\hat{A}$ will be estimated as a nearly zero matrix. Denote $\boldsymbol{X_i}$ by one of the other exogenous variables. We obtain

$$\begin{aligned}
&\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} \\
&= \log\tau_{m_j}\left(\hat{A}\hat{\Sigma}_{\Psi(\boldsymbol{X_j})}\hat{A}^T\right) - \log\tau_{m_j}\left(\hat{A}\hat{A}^T\right) - \log\tau_N\left(\hat{\Sigma}_{\Psi(\boldsymbol{X_j})}\right) \\
&= \log\left[\frac{1}{N}\tau_{m_j}\left(\hat{A}\boldsymbol{G_j}\hat{A}^T\right)\right] - \log\tau_{m_j}\left(\hat{A}\hat{A}^T\right) - \log\left[\frac{1}{N}\tau_N\left(\boldsymbol{G_j}\right)\right] \\
&= \log\tau_{m_j}\left(\hat{A}\boldsymbol{G_j}\hat{A}^T\right) - \log\tau_{m_j}\left(\hat{A}\hat{A}^T\right) - \log\tau_N\left(\boldsymbol{G_j}\right). \qquad (9)
\end{aligned}$$

Since $\hat{A}$ is an estimated nearly zero matrix, $\log\tau_{m_j}(\hat{A}\boldsymbol{G_j}\hat{A}^T) \approx \log\tau_{m_j}(\hat{A}\hat{A}^T)$ holds. Due to the definition of kernel Gram matrix and τ_N, we have $\tau_N(\boldsymbol{G_j}) = 1$, resulting in $\log\tau_N(\boldsymbol{G_j}) = 0$. Thus, it induces $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} \approx 0$. Second, in the cases where denote $\boldsymbol{X_k}$ by one of the other exogenous variables and $\boldsymbol{X_i}$ one of the nonexogenous variables, if $\boldsymbol{X_j}$ and $\boldsymbol{X_i}$ share deterministic relationships as in case (a-1) of Fig. 2 or they are independent as in case (a-2), we see that $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} \approx 0$ holds. As in case (a-3), if $\boldsymbol{X_j}$ and $\boldsymbol{X_i}$ do not satisfy the deterministic relationships, thereafter interestingly, with the transformation $\boldsymbol{X_{\overline{j}}} = [\boldsymbol{X_j}, \boldsymbol{X_k}]^T$, we see that $\boldsymbol{X_i} = \boldsymbol{f_{ik}}(\boldsymbol{X_k}) + \boldsymbol{f_{ij}}(\boldsymbol{X_j}) = \hat{\boldsymbol{f}}(\boldsymbol{X_{\overline{j}}})$, which means that $\boldsymbol{X_{\overline{j}}}$ and $\boldsymbol{X_i}$ are deterministic and $\Delta_{\Psi(\boldsymbol{X_{\overline{j}}})\to\boldsymbol{X_i}} \approx 0$ holds. Hence, if $\boldsymbol{X_j}$ is exogenous, we prove that $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} \approx 0$ holds for all $i \neq j$ holds.

Second, assume that $\boldsymbol{X_j}$ is not exogenous, and we prove that there at least exists one variable satisfying $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} < 0$. Since $\boldsymbol{X_j}$ is not exogenous, $\boldsymbol{X_j}$ has at least one parent. If $\boldsymbol{X_j}$ has only one parent, denote this parent by $\boldsymbol{X_i}$. In this case, $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ directly satisfy the deterministic relationships,[2] as shown in case (b-1) of Fig. 2, then according to (4) of the KTR method, $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} < 0$ holds. If $\boldsymbol{X_j}$ has more than one parents and we assume that it has two parents $\boldsymbol{X_i}$ and $\boldsymbol{X_k}$, where possible structures are shown in cases (b-2) and (b-3) of Fig. 2, then $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ do not directly satisfy the deterministic relationships. In case (b-2) of Fig. 2, through transforming $\boldsymbol{X_i}$ into $\boldsymbol{X_{\overline{i}}} = [\boldsymbol{X_i}, \boldsymbol{X_k}]^T$, i.e., $\boldsymbol{X_j} = \boldsymbol{f_{ji}}(\boldsymbol{X_i}) + \boldsymbol{f_{jk}}(\boldsymbol{X_k}) = \hat{\boldsymbol{f}}(\boldsymbol{X_{\overline{i}}})$, we see that $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_{\overline{i}}}} < 0$ holds. In case (b-3), $\boldsymbol{X_j}$ and $\boldsymbol{X_k}$ using model (1) can still follow the deterministic relationships, $\boldsymbol{X_j} = \boldsymbol{f_{ji}}(\boldsymbol{X_i}) + \boldsymbol{f_{jk}}(\boldsymbol{X_k}) = \hat{\boldsymbol{f}}(\boldsymbol{X_k})$, which implies that $\Delta_{\Psi(\boldsymbol{X_k})\to\boldsymbol{X_j}} \approx 0$ and $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_k}} < 0$. Thus, it is proven that if $\boldsymbol{X_j}$ is not exogenous, there always exists at least one variable $\boldsymbol{X_i}$ or $\boldsymbol{X_k}$ such that $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} < 0$ or $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_k}} < 0$ holds. To be concluded, the theorem is proven. □

Theorem 2 indicates the property of an exogenous variable that can be utilized for identification. Furthermore, we need to demonstrate the validity of iteratively selecting the exogenous variables, i.e., the consistency of HDDM.

Theorem 3: Assume that data $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ strictly follow the HDDM (1) and variable $\boldsymbol{X_k}$ is exogenous. Then, when identifying the next exogenous variable, any two dependent variables $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ with $i, j \neq k, i \neq j$ and $K(j) < K(i)$ still have deterministic relationships through transforming $\boldsymbol{X_j}$ to $\boldsymbol{X_{\overline{j}}}$, i.e., $\boldsymbol{X_i} = \hat{\boldsymbol{f}}(\boldsymbol{X_{\overline{j}}})$, where $\hat{\boldsymbol{f}}$ are injective and nonlinear (or linear) functions. $\boldsymbol{X_{\overline{j}}}$ is a union variable that concatenates the parental and ancestral variables of $\boldsymbol{X_i}$, including $\boldsymbol{X_j}$, and the variable $\boldsymbol{X_k}$.

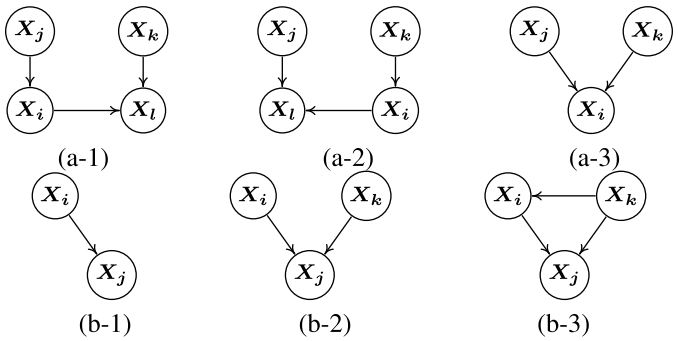

Fig. 2. Local illustrative graphs for the proof of Theorem 2. (a-1)–(a-3) Possible relationships between $\boldsymbol{X_j}$ and $\boldsymbol{X_i}$ when there is more than one exogenous variable in the graph (assuming that $\boldsymbol{X_j}$ and $\boldsymbol{X_k}$ are exogenous) and $\boldsymbol{X_k}$ is nonexogenous. (b-1)–(b-3) Possible local structures when $\boldsymbol{X_j}$ is not exogenous.

In other words, Theorem 3 guarantees that the HDDM still possesses the characteristics of deterministic relationships for any dependent pairs in the data after identifying the first exogenous variable.

Proof: Without loss of generality, since that data $\boldsymbol{D}$ follow the HDDM, we have $\boldsymbol{X_i} = \hat{\boldsymbol{f}}(\boldsymbol{X_{\overline{j}}})$, where $i, j \in \{1, \ldots, n\}$ and $i \neq j$. Note that if $\boldsymbol{X_k}$ is selected as an exogenous variable, i.e., the causal ordering of $\boldsymbol{X_k}$ termed $K(k)$ will be earlier than other variables, we will take the use of this ordering information priorly without igoring $\boldsymbol{X_k}$. Concretely, we would add $\boldsymbol{X_k}$ into every transformed variable as demonstrated in Rule 2.

Due to $K(j) < K(i)$, we prove the theorem from two aspects. First, $\boldsymbol{X_i}$ contains the information of $\boldsymbol{X_k}$, that is, $\boldsymbol{X_k}$ directly or indirectly points to $\boldsymbol{X_i}$. Hence, when identifying the next exogenous, $\boldsymbol{X_i} = \hat{\boldsymbol{f}}(\boldsymbol{X_{\overline{j}}})$ still holds since $\boldsymbol{X_k}$ is also included by the union variable $\boldsymbol{X_{\overline{j}}}$. Second, $\boldsymbol{X_i}$ does not contain the information of $\boldsymbol{X_k}$. In this case, when identifying the first exogenous variable, $\boldsymbol{X_i} = \hat{\boldsymbol{f}}(\boldsymbol{X_{\overline{j}}})$ holds, where $\boldsymbol{X_k}$ is not included in $\boldsymbol{X_{\overline{j}}}$ because $\boldsymbol{X_k}$ is not a parental or ancestral variable of $\boldsymbol{X_i}$. When identifying the next, we would add $\boldsymbol{X_k}$ into $\boldsymbol{X_{\overline{j}}}$, i.e.,

$$\begin{aligned}
\boldsymbol{X_i} &= \sum_{K(j)<K(i)} \boldsymbol{f_{ij}}(\boldsymbol{X_j}) = \boldsymbol{f_{ij}}(\boldsymbol{X_j}) + \boldsymbol{f_{il}}(\boldsymbol{X_l}) \\
&= \boldsymbol{f_{ij}}\left(\boldsymbol{A_1}\cdot\boldsymbol{X_{\overline{j}}}\right) + \boldsymbol{f_{il}}\left(\boldsymbol{A_2}\cdot\boldsymbol{X_{\overline{j}}}\right) \\
&= \hat{\boldsymbol{f}}\left(\boldsymbol{X_{\overline{j}}}\right) \qquad (10)
\end{aligned}$$

where we assume that $\boldsymbol{X_l}$ is another variable that supports $K(l) < K(i), l \neq i, j, k$, $\boldsymbol{X_{\overline{j}}} = [\boldsymbol{X_j}, \boldsymbol{X_l}, \boldsymbol{X_k}]^T$ is the newly union variable transformed by $\boldsymbol{X_j}$, and $\boldsymbol{A_1} = [\boldsymbol{I}_{m_j\times m_j}, \boldsymbol{0}_{m_j\times m_l}, \boldsymbol{0}_{m_j\times m_k}]$ and $\boldsymbol{A_2} = [\boldsymbol{0}_{m_l\times m_j}, \boldsymbol{I}_{m_l\times m_l}, \boldsymbol{0}_{m_l\times m_k}]$ are two coefficient matrices. It is noted that when added $\boldsymbol{X_k}$ into $\boldsymbol{X_{\overline{j}}}$, any pair of $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ still satisfies the deterministic relationships. To be concluded, the theorem is proven. □

[2]We say here that $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ directly satisfy the deterministic relationships if there is only one directed path from $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ and the indegree of $\boldsymbol{X_j}$ is equal to 1.

Algorithm 1 HDDM Algorithm

1: Input: Dataset $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ where the variable $\boldsymbol{X_i}$ is m_i-dimensional, sample size N, regularization parameter λ, and decision parameter α.
2: Output: A causal ordering set K.
3: Initialize an unordered set $U = \{1, \ldots, n\}$ and a causal ordering set $K = \varnothing$;
//Map the data into Kernel space
4: **for** $i = 1$ to n **do**
5: Normalize $\boldsymbol{X_i}$ for each dimension;
6: Compute $\boldsymbol{X_i}$'s kernel Gram matrix $\boldsymbol{G_i}$;
7: **end for**
//Estimate a causal ordering
8: **while** $U \neq \varnothing$ **do**
9: **if** $K = \varnothing$ **then**
10: Identify the first exogenous variable $\boldsymbol{X_k}$ using Algorithm 2;
11: Append the index k to K;
12: $U = U \setminus k$;
13: **else**
14: Identify the next exogenous variable $\boldsymbol{X_k}$ using Algorithm 3;
15: Append the index k to K;
16: $U = U \setminus k$.
17: **end if**
18: **end while**

Theorems 2 and 3 assure the correctness and validate the consistency for our proposed method theoretically. They also show the rationality of the HDDM model.

C. Algorithm

Based on the theoretical results in Section IV-B, we propose Algorithm 1, called HDDM algorithm to estimate the causal ordering among multiple high-dimensional variables. We then give two examples to better illustrate the algorithm.

As demonstrated, we first compute each variable's Gram matrix using the empirical kernel mapping. This aims at tackling the nonlinear relationships between variables. Then, we select the exogenous variable iteratively until the whole causal ordering is determined. Specifically, since there is no prior information when we identify the first exogenous variable, we would use Rule 1 for the first identification, and for the next, we would use Rule 2, which are shown in Algorithms 2 and 3, respectively.

In Algorithm 2, as shown in Rule 1, we first compute the $\boldsymbol{K}_{ij}$ for every pair $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ in the data.[3] If this pair directly satisfies deterministic relationships or is independent with each other, we will mark them in the decision matrix $\boldsymbol{A}$ in lines 8, 10, or 12. Otherwise, if they do not satisfy, we will transform $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ that concatenate all variables except $\boldsymbol{X_j}$ and $\boldsymbol{X_i}$, respectively, and we further mark the satisfying pairs in $\boldsymbol{A}$. Eventually, according to Theorem 2, we compute the summation of each row m_i of $\boldsymbol{A}$ and choose the one with a maximum m_i as the first root variable.

[3]For the sake of simplicity, in Algorithms 2 and 3, we define a KTR matrix $\boldsymbol{K}$ that stores all the scale-invariant measures, i.e., $\boldsymbol{K}_{ji} = \Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}}$.

Algorithm 2 Identifying the First Exogenous Variable

1: Input: Dataset $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ where the variable $\boldsymbol{X_i}$ is m_i-dimensional, regularization parameter λ, and decision parameter α.
2: Output: The first root variable $\boldsymbol{X_k}$, the kernelized trace matrix $\boldsymbol{K}$, the decision matrix $\boldsymbol{A}$.
//Using **Rule 1**
3: **for** $i = 1$ to $(n-1)$ **do**
4: **for** $j = (i+1)$ to n **do**
5: $\boldsymbol{V_i} = \boldsymbol{X_i}, \boldsymbol{V_j} = \boldsymbol{X_j}$
6: Compute $\boldsymbol{K}_{ij}$ and $\boldsymbol{K}_{ji}$, using $\boldsymbol{V_i}, \boldsymbol{V_j}$ and Eq.(6);
7: **if** $\boldsymbol{K}_{ij} \approx 0$ and $|\boldsymbol{K}_{ij} - \boldsymbol{K}_{ji}| > \alpha$ **then**
8: $\boldsymbol{A}_{ij} = 1, \boldsymbol{A}_{ji} = 0$;
9: **else if** $\boldsymbol{K}_{ji} \approx 0$ and $|\boldsymbol{K}_{ij} - \boldsymbol{K}_{ji}| > \alpha$ **then**
10: $\boldsymbol{A}_{ji} = 1, \boldsymbol{A}_{ij} = 0$;
11: **else if** $\boldsymbol{K}_{ij} \approx 0$ and $\boldsymbol{K}_{ji} \approx 0$ **then**
12: $\boldsymbol{A}_{ij} = 1, \boldsymbol{A}_{ji} = 1$;
13: **end if**
14: **if** $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji} = 0$ **then**
15: $\boldsymbol{V_i^1} = \boldsymbol{D} \setminus \boldsymbol{X_j}, \boldsymbol{V_j^1} = \boldsymbol{X_j}$;
16: Compute $\boldsymbol{K}_{ij}$, using $\boldsymbol{V_i^1}, \boldsymbol{V_j^1}$ and Eq.(6);
17: $\boldsymbol{V_i^2} = \boldsymbol{X_i}, \boldsymbol{V_j^2} = \boldsymbol{D} \setminus \boldsymbol{X_i}$;
18: Compute $\boldsymbol{K}_{ji}$, using $\boldsymbol{V_i^2}, \boldsymbol{V_j^2}$ and Eq.(6);
19: **if** $\boldsymbol{K}_{ij} \approx 0$ and $|\boldsymbol{K}_{ij} - \boldsymbol{K}_{ji}| > \alpha$ **then**
20: $\boldsymbol{A}_{ij} = 1, \boldsymbol{A}_{ji} = 0$;
21: **else if** $\boldsymbol{K}_{ji} \approx 0$ and $|\boldsymbol{K}_{ij} - \boldsymbol{K}_{ji}| > \alpha$ **then**
22: $\boldsymbol{A}_{ji} = 1, \boldsymbol{A}_{ij} = 0$;
23: **end if**
24: **end if**
25: **end for**
26: **end for**
27: Compute $m_i = \sum_{j \in U \setminus i} \boldsymbol{A}_{ij}$;
28: Select the index of the root variable: $k = \arg\max_{i \in U} m_i$.

In Algorithm 3, to identify the next root, we utilize Rule 2, which is guaranteed to be feasible by Theorem 3. As shown, we have prior information from the causal ordering set K. Hence, every time before we compute the quantities of interest $\hat{\boldsymbol{K}}_{ij}$ and $\hat{\boldsymbol{K}}_{ji}$, we will make a transformation toward $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$, respectively. Note that in lines 9–12, there are two conditions to be met to update the decision matrix $\boldsymbol{A}$: 1) $\hat{\boldsymbol{K}}_{ij}$ itself is equal to nearly zero and 2) the newly difference between $\hat{\boldsymbol{K}}_{ij}$ and $\hat{\boldsymbol{K}}_{ji}$ is larger than the difference between the input $\boldsymbol{K}_{ij}$ and $\boldsymbol{K}_{ji}$, i.e., $|\hat{\boldsymbol{K}}_{ij} - \hat{\boldsymbol{K}}_{ji}| > |\boldsymbol{K}_{ij} - \boldsymbol{K}_{ji}|$. This comes from the property that satisfying these two conditions implies a causal relation from the picked variables, combined with the causal ordering set K, to the effect variable. Finally, we compute the summation of each row m_i of $\boldsymbol{A}$ and choose the one with a maximum m_i as the next root variable, which is the same as Algorithm 2.

In the following, we give two toy examples for better illustration for our proposed HDDM algorithm, i.e., how to employ Rules 1 and 2 based on the theoretical results.

Example 1: As shown in Fig. 3(a), consider a causal structure where there are four high-dimensional variables $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, and $\boldsymbol{X_4}$. It is worth noting that if two variables $\boldsymbol{X_i}$ and $\boldsymbol{X_j}$ satisfy a deterministic relationship, $\Delta_{\Psi(\boldsymbol{X_j})\to\boldsymbol{X_i}} \approx 0$ in (3)

Algorithm 3 Identifying the Next Exogenous Variable

Input: Dataset $\boldsymbol{D} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_n}]^T$ where the variable $\boldsymbol{X_i}$ is m_i-dimensional, regularization parameter λ, kernelized trace matrix $\boldsymbol{K}$, an unordered set U, a causal ordering set K, and the decision matrix $\boldsymbol{A}$.
Output: The next exogenous variable $\boldsymbol{X_k}$, the kernelized trace matrix $\boldsymbol{K}$, the decision matrix $\boldsymbol{A}$.
//Using **Rule 2**
for $i \in U$ **do**
  for $j \in U \setminus i$ **do**
    $\boldsymbol{U_i} = \boldsymbol{X_i} \cup \boldsymbol{X}_K, \boldsymbol{U_j} = \boldsymbol{X_j}$;
    Compute $\hat{\boldsymbol{K}}_{ij}$, using $\boldsymbol{U_i}, \boldsymbol{U_j}$ and Eq.(6);
    $\boldsymbol{U_i} = \boldsymbol{X_i}, \boldsymbol{U_j} = \boldsymbol{X_j} \cup \boldsymbol{X}_K$;
    Compute $\hat{\boldsymbol{K}}_{ji}$, using $\boldsymbol{U_i}, \boldsymbol{U_j}$ and Eq.(6);
    if $\hat{\boldsymbol{K}}_{ij} \approx 0$ and $|\hat{\boldsymbol{K}}_{ij} - \hat{\boldsymbol{K}}_{ji}| > |\boldsymbol{K}_{ij} - \boldsymbol{K}_{ji}|$ **then**
      $\boldsymbol{A}_{ij} = 1, \boldsymbol{A}_{ji} = 0$;
    else if $\hat{\boldsymbol{K}}_{ji} \approx 0$ and $|\hat{\boldsymbol{K}}_{ij} - \hat{\boldsymbol{K}}_{ji}| > |\boldsymbol{K}_{ij} - \boldsymbol{K}_{ji}|$ **then**
      $\boldsymbol{A}_{ji} = 1, \boldsymbol{A}_{ij} = 0$;
    else if $\hat{\boldsymbol{K}}_{ij} \approx 0$ and $\hat{\boldsymbol{K}}_{ji} \approx 0$ **then**
      $\boldsymbol{A}_{ij} = 1, \boldsymbol{A}_{ji} = 1$;
    end if
  end for
end for
Compute $m_i = \sum_{j \in \boldsymbol{U} \setminus i} A_{ij}$;
Select the index of the next root variable: $k = \arg\max_{i \in \boldsymbol{U}} m_i$.

will hold for the causal direction, while it will be violated for the anticausal direction, i.e., $\Delta_{\Psi(\boldsymbol{X_i}) \to \boldsymbol{X_j}} \leq 0$. Identically, if $\Delta_{\Psi(\boldsymbol{X_j}) \to \boldsymbol{X_i}} \approx 0$ or $\Delta_{\Psi(\boldsymbol{X_i}) \to \boldsymbol{X_j}} \leq 0$ does not hold, then these two variables do not share the deterministic relationship. To identify the first exogenous variable, we employ Rule 1 in Algorithm 2. We find that pairs $\{\boldsymbol{X_1}, \boldsymbol{X_2}\}$, $\{\boldsymbol{X_1}, \boldsymbol{X_3}\}$, and $\{\boldsymbol{X_1}, \boldsymbol{X_4}\}$ satisfy deterministic relationships without transformations, so we mark the decision matrix $\mathbf{A}_{12} = 1$, $\mathbf{A}_{13} = 1$, and $\mathbf{A}_{14} = 1$. Since the pairs $\{\boldsymbol{X_2}, \boldsymbol{X_3}\}$, $\{\boldsymbol{X_2}, \boldsymbol{X_4}\}$, and $\{\boldsymbol{X_3}, \boldsymbol{X_4}\}$ do not own this relationships, we would make transformations toward every variable of them. For instance, for the pair $\{\boldsymbol{X_2}, \boldsymbol{X_4}\}$, we transform $\boldsymbol{X_2}$ and $\boldsymbol{X_4}$ into $\boldsymbol{X}_{\overline{\boldsymbol{2}}} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}]^T$ and $\boldsymbol{X}_{\overline{\boldsymbol{4}}} = [\boldsymbol{X_1}, \boldsymbol{X_3}, \boldsymbol{X_4}]^T$ and subsequently compute its statistics of interest of $\{\boldsymbol{X}_{\overline{\boldsymbol{2}}}, \boldsymbol{X_4}\}$ and $\{\boldsymbol{X}_{\overline{\boldsymbol{4}}}, \boldsymbol{X_2}\}$, respectively. It is found that $\{\boldsymbol{X}_{\overline{\boldsymbol{2}}}, \boldsymbol{X_4}\}$ satisfies the deterministic relationship, so we mark $\mathbf{A}_{24} = 1$. Similar procedures are applied to the pairs $\{\boldsymbol{X_2}, \boldsymbol{X_3}\}$ and $\{\boldsymbol{X_3}, \boldsymbol{X_4}\}$, resulting in $\mathbf{A}_{34} = 1$. According to line 27 of Algorithm 2, $\boldsymbol{X_1}$ is selected as the first exogenous variable. Once we obtain the information of causal ordering, we would utilize Rule 2 in Algorithm 3 to identify the next exogenous. Concretely, we compute the interests for the pairs $\{\boldsymbol{X_2}, \boldsymbol{X_3}\}$, $\{\boldsymbol{X_2}, \boldsymbol{X_4}\}$, and $\{\boldsymbol{X_3}, \boldsymbol{X_4}\}$. For instance, for the pair $\{\boldsymbol{X_2}, \boldsymbol{X_3}\}$, we compute the interests of $\{\boldsymbol{X}_{\overline{\boldsymbol{2}}}, \boldsymbol{X_3}\}$ and $\{\boldsymbol{X}_{\overline{\boldsymbol{3}}}, \boldsymbol{X_2}\}$, where $\boldsymbol{X}_{\overline{\boldsymbol{2}}} = [\boldsymbol{X_1}, \boldsymbol{X_2}]^T$ and $\boldsymbol{X}_{\overline{\boldsymbol{3}}} = [\boldsymbol{X_1}, \boldsymbol{X_3}]^T$. We found that $\{\boldsymbol{X}_{\overline{\boldsymbol{2}}}, \boldsymbol{X_3}\}$ satisfies the deterministic relationship so we mark $\mathbf{A}_{23} = 1$. Applying similar procedures for other pairs, we have $\mathbf{A}_{24} = 1$ and $\mathbf{A}_{34} = 1$, so $\boldsymbol{X_2}$ is selected as the next exogenous variable. Employing Algorithm 3 again, we obtain $\boldsymbol{X_3}$ as the next exogenous variable. Eventually, the whole ordering can be determined.

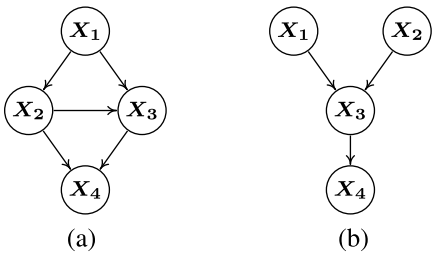

Fig. 3. Examples of causal structures for better interpretation of our proposed algorithm. (a) Example 1 where there are 4 high-dimensional variables with 5 edges. (b) Example 2 where there are 4 high-dimensional variables with 3 edges.

Example 2: As shown in Fig. 3(b), to estimate the whole causal ordering, we perform the HDDM method in Algorithm 1. We first find that the pair $\{\boldsymbol{X_3}, \boldsymbol{X_4}\}$ follows the deterministic relationships, so the decision matrix is marked as $\mathbf{A}_{34} = 1$. Since $\boldsymbol{X_1}$ is independent with $\boldsymbol{X_2}$, we have $\boldsymbol{K}_{ij} \approx 0$ and $\boldsymbol{K}_{ji} \approx 0$, resulting in $\mathbf{A}_{12} = 1$ and $\mathbf{A}_{21} = 1$. Then, we compute the statistics of interests for the pairs $\{\boldsymbol{X_1}, \boldsymbol{X_3}\}$, $\{\boldsymbol{X_1}, \boldsymbol{X_4}\}$, $\{\boldsymbol{X_2}, \boldsymbol{X_3}\}$, and $\{\boldsymbol{X_2}, \boldsymbol{X_4}\}$ after the transformation procedures and have $\mathbf{A}_{14} = 1$ and $\mathbf{A}_{24} = 1$. According to line 27 of Algorithm 2, suppose that $\boldsymbol{X_2}$ is selected as the first exogenous variable (selecting $\boldsymbol{X_1}$ as the first is identical). With Algorithm 3, we acquire $\mathbf{A}_{13} = 1$, $\mathbf{A}_{14} = 1$, and $\mathbf{A}_{34} = 1$, from which $\boldsymbol{X_1}$ is selected as the next exogenous. Utilizing similar procedures, we finally determine the whole causal ordering.

V. Experiments

In this section, we conducted experiments on synthetic data based on various real networks and real-world data to verify the effectiveness of our proposed method. We tested the sensitivity to dimensions of each variable in Section V-A, the sensitivity to samples with distinct real networks[4] in Section V-B, the sensitivity to noises in Section V-C, and the sensitivity to parameters in Section V-D, and we experimented real data in Section V-E.

We compared our method with the state-of-the-art methods, including FI [20] and GDL [18], where the latter algorithm has four variants, i.e., GDL using nonlinear correlation or Hilbert schmidt independence criterion (HSIC) as independence tests (abbreviated as GDL-nlcorr and GDL-HSIC respectively), trace method (abbreviated as TrMeth), and pairwise measure (abbreviated as PwMeas) [18]. Since FI has outperformed EMD in the case of multiple nonlinear variables, we do not make a comparison with EMD here. Besides, we take the accuracies of identifying the first exogenous variables and the whole causal ordering as our measurement criteria.

Unless specified, the detailed default experimental settings are used in each section, which are described in the following. The default real network for synthetic data is called *survey*, which has six variables. Each generated variable is 5-D. We fix the sample size as 500. To construct the first exogenous variable denoted as $\boldsymbol{X_j}$, we use the same way as [16]: we first generate $Z = Cd$, where $C_{ij} \sim \mathcal{N}(0, 1)$ and $d_i \sim \mathcal{N}(0, 1)$. Then, we pass Z_i through its cumulative distribution function

[4]The real networks are Bayesian networks from http://www.bnlearn.com/bnrepository/

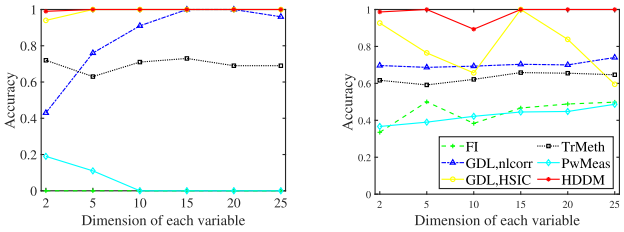

Fig. 4. Causal ordering identification performances with different dimensions for each variable. (a) First exogenous variable. (b) Whole ordering.

to transform Z_i into uniform distribution, i.e., $X_{ji} = F_{Z_i}(Z_i)$, where X_{ji} represents the ith dimension of $\boldsymbol{X_j}$. Finally, X_{ji} is rescaled and shifted to $[-2, 2]$. Through this setting, $p(\boldsymbol{X_j})$ follows a uniform distribution on $[-2, 2]$ and the dependencies among X_{ji} are constructed. To better verify our method's efficacy, we use general data generation mechanisms $\boldsymbol{X_i} = f(\boldsymbol{X_{Pa_i}})$, where $\boldsymbol{X_{Pa_i}}$ are the parents of $\boldsymbol{X_i}$, to element-wisely construct other high-dimensional variables. We use different nonlinear functions for different numbers of parents, i.e., $f(\boldsymbol{X_j}) = \text{sign}(\boldsymbol{X_j})|\boldsymbol{X_j}|^2$, $f(\boldsymbol{X_j}, \boldsymbol{X_k}) = |\boldsymbol{X_j}|^2 \cos(\boldsymbol{X_k})$, or $f(\boldsymbol{X_j}, \boldsymbol{X_k}, \boldsymbol{X_l}) = \text{sign}(\boldsymbol{X_j})|\boldsymbol{X_k}|^2 \cos(\boldsymbol{X_l})$. Note that in this way, it is reasonable to assume that the uniform distributions and the nonlinear function are independently generated [16]. We use the Gaussian kernel to compute every variable's Gram matrix with the kernel width fixed to be $2/9$ of the median distance. The regularization parameter λ is set to be $1e^{-5}$. Such two parameters could also be determined using the n-fold cross-validation method.

A. Sensitivity to Dimensions

In this section, we tested the performance of the algorithms on various dimensions of each variable. These dimensions were taken from $\{2, 5, 10, 15, 20, 25\}$. We randomly generated the data using the aforementioned nonlinear functions and according to a real network called survey, which has six variables. Note that it was not generated from our HDDM in (1) but a more general data mechanism since this general generation method is more convincing and fairer. For each dimensionality in $\{2, 5, 10, 15, 20, 25\}$, we independently conducted 100 trials and computed its average accuracy of finding the first exogenous variable and the whole ordering with results shown in Fig. 4. Our proposed method HDDM outperforms other methods, especially in the identification accuracy of the first variable. It indicates that the candidates' selection rules did help when estimating the causal ordering under different dimensions of each variable. Furthermore, it also confirms the correctness of our theoretical analysis. FI has an unexpectedly unsatisfactory performance compared with other methods, and the reason might be that it cannot guarantee that pairs satisfy the model assumptions.

B. Sensitivity to Samples With Distinct Real Networks

This section aimed at verifying the performance under different sample sizes $\{100, 200, 300, 400, 500\}$ and various real networks[4] *cancer*, *survey*, *asia* and *sachs*, where the number of variables of these networks are 5, 6, 8, and 11, respectively. With each sample size of each real network, we also sampled 100 times and computed the accuracies. Data were generated in the same way as Section V-A, using the general generation mechanism under the real networks. Each variable is 5-dimensional. Results are illustrated in Fig. 5. It can be seen that our method outperforms other methods in almost all cases. In particular, as for small networks such as cancer and survey, HDDM obtains 100% accuracy in estimating the first exogenous variable as well as the whole ordering. As for other two larger networks asia and sachs, when the sample size is 400 or 500, HDDM does not identify the first exogenous variable well, and the GDL-nlcorr method has a similar performance. This may result from the unstableness of estimated quantities. However, for the whole causal ordering, there is a slight improvement for most methods as the sample size increases. Among all, HDDM performs the best. It also verifies the effectiveness of the proposed rules with different sample sizes and distinct real networks, especially in small networks.

C. Sensitivity to Noises

Since our method is focused on deterministic models, in this section, we would illustrate the degree to which identifiability is hampered by noise. The noise ratios change from 0.1 to 1.0. The data generation process is identical to the previous experimental sections. The default real network is survey, the number of variables is 6, and each variable is 5-D. After conducting 100 independent trials with 500 samples, we present the results in Fig. 6. As can be seen, the identification performance of the first exogenous variable for HDDM remains stable as 100% accuracy as noise ratios increase, while other methods are not. This shows the potential power of Rule 1 to handle undeterministic cases. To estimate the whole ordering, the HDDM performance raises to 100% as the noise ratio goes up to 0.6. Although it lacks justification in the presence of additive noises, HDDM works empirically well. Other methods, including TrMeth and PwMeas, are stable toward the noises, and however, their performances are not satisfactory enough. This is because the nonlinear functions violate their assumptions.

D. Sensitivity to Parameters

Two parameters are involved with our algorithm, kernel width and regularization parameter λ. In this section, we examined the robustness of our method toward these two parameters. We also utilized the same mechanism to generate the data as Section V-A, with the six-variable real network survey, and the dimensionality of each variable was 5. The sample size was set to be 500. For examining the robustness of the kernel widths, we first fixed the regularization parameter λ to be $1e^{-5}$ and set kernel widths to be different values, i.e., 2/10, 2/9, 2/8, 2/7, 2/6, 2/5, 2/4, 2/3, 1, 2, 4, 6, 8, and 10 of the median distance. For examining the robustness of the regularization parameter λ, we first set the kernel width to be 2/9 of the median distance and changed λ to be 1, $1e^{-1}$, $1e^{-2}$, $1e^{-3}$, $1e^{-4}$, $1e^{-5}$, $1e^{-6}$, $1e^{-7}$, $1e^{-8}$, $1e^{-9}$, and $1e^{-10}$. The corresponding performance is shown in Fig. 7.

As shown in Fig. 7, the whole ordering performance drops as the kernel width is 2/8 and keeps stable as it is larger

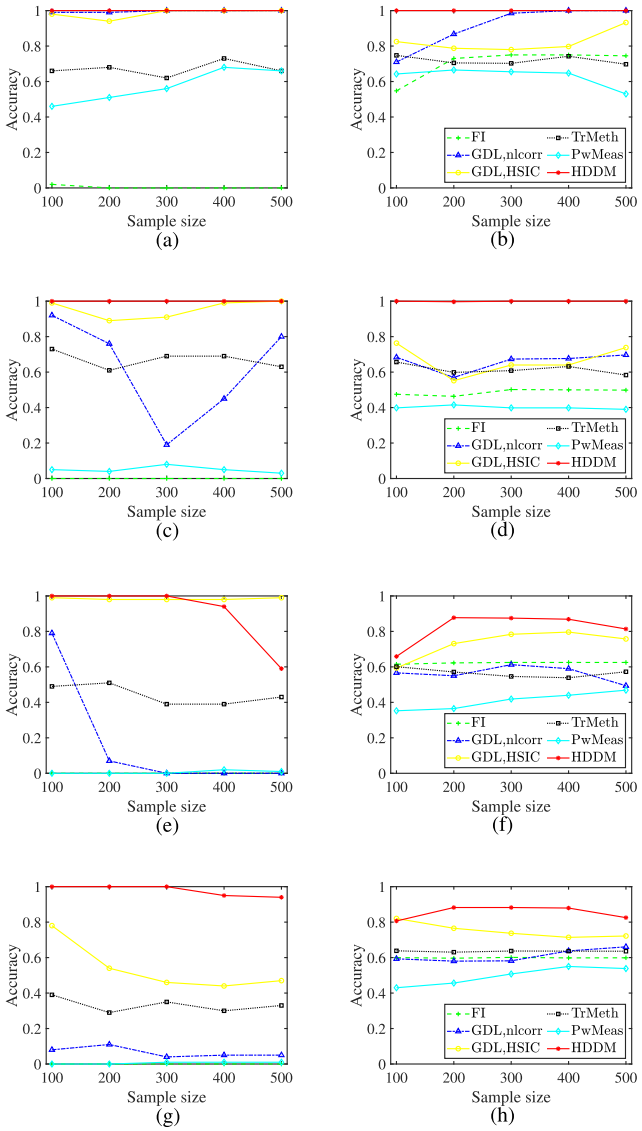

Fig. 5. Causal ordering identification performances with different real networks and different sample sizes. (a) First exogenous variable of *cancer*. (b) Whole ordering of *cancer*. (c) First exogenous variable of *survey*. (d) Whole ordering of *survey*. (e) First exogenous variable of *asia*. (f) Whole ordering of *asia*. (g) First exogenous variable of *sachs*. (h) Whole ordering of *sachs*.

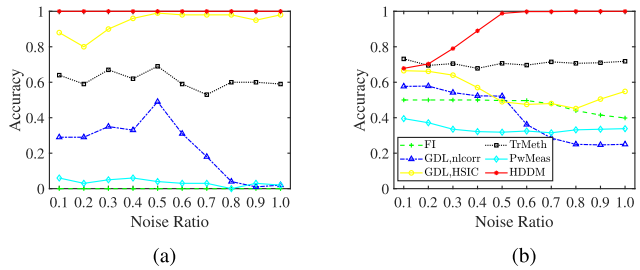

Fig. 6. Causal ordering identification performances with different noise ratios. (a) First exogenous variable. (b) Whole ordering.

than 2/6. When it is equal to 2/9, our method obtains 100% identification accuracy. As for the regularization parameter λ, when λ is large, the performance is not such satisfactory. This is because errors will be increased in the estimated $\hat{A}$ in (6) if λ is large, as a result of low identification accuracy of the whole causal ordering. When λ is too small, the performance shows a deterioration, which may result from the model overfit of the data. Thus, it is nontrivial to select the appropriate parameters for estimation.

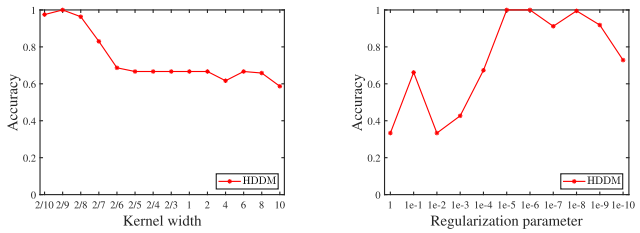

Fig. 7. Whole causal ordering identification performances with different kernel widths and regularization parameters. (a) Kernel widths. (b) Regularization parameters.

TABLE I

CAUSAL ORDERING IDENTIFICATION PERFORMANCES OF STOCK INDICES DATA, WHERE THE GROUND TRUTH IS Asia → Europe → America DUE TO THE TIME DIFFERENCE AND THE EFFICIENT MARKET HYPOTHESIS. THE CAUSAL ORDERING THEN IS {Asia, Europe, America} [11], [16], [19]

Algorithms	Causal orderings
Ground truth	{Asia, Europe, America}
FI	{Asia, America, Europe}
GDL-nlcorr	{America, Europe, Asia}
GDL-HSIC	{America, Europe, Asia}
TrMeth	{Europe, Asia, America}
PwMeas	{America, Europe, Asia}
HDDM (Ours)	{Asia, Europe, America}

E. Real Data

1) Yahoo Finance Data: We aimed at discovering a causal ordering in the daily stock returns between different regions of the world, i.e., America, Europe, and Asia, where each region consisted of two or four stock indices in its area. They were America := {BSESN, DJI, GSPC, IXIC} from the USA, Europe := {BUK100P, FCHI} from Germany and France, and Asia := {HSI, N225} from China and Japan. The data were downloaded from the Yahoo finance database at 1041 days (from February 10, 2015, to January 10, 2020) and we used the adjusted closing prices for the stocks. We normalized each stock so that they have zero means and unit variances. Note that here we did not test whether such data satisfy the deterministic causal relationship or not, since our method as shown in the synthetic experiments could work empirically well even in the presence of noises.

We compared our method HDDM with other existing methods, i.e., FI, GDL-nlcorr, GDL-HSIC, TrMeth, and PwMeas, and applied the same settings as those applied in the synthetic data. Results are shown in Table I. Due to the time difference and the efficient market hypothesis, it was assumed that there existed a ground truth of the causal ordering [11], [16], [19]: Asia → Europe → America. From Table I, we found that HDDM obtained a better causal ordering among different regions with stock indices, which was in accordance with the ground truth. The results of GDL-nlcorr, GDL-HSIC, and PwMeas implied that the data generation of this stock indices data might not be simply linear, so they cannot correctly identify the orderings. Compared with FI and TrMeth, the superior

TABLE II

CAUSAL ORDERING IDENTIFICATION PERFORMANCES OF FMRI HIPPOCAMPUS DATA, WHERE THE GROUND TRUTH IS PRC → ERC, PHC → ERC, ERC → CA3/DG, ERC → CA1, CA3/DG → CA1, CA1 → SUB, AND SUB → ERC [47], [48]. THUS, THE CAUSAL ORDERING CAN BE ASSUMED TO BE {PRC, PHC, ERC, CA3/DG, CA1, AND SUB}

Algorithms	Causal orderings
Ground truth	{ PRC, PHC, ERC, CA3/DG, CA1, Sub }
FI	{ CA3/DG, PRC, ERC, PHC, Sub, CA1 }
GDL-nlcorr	{ CA1, Sub, CA3/DG, PHC, PRC, ERC }
GDL-HSIC	{ PRC, PHC, ERC, CA3/DG, CA1, Sub }
TrMeth	{ Sub, ERC, PRC, PHC, CA3/DG, CA1 }
PwMeas	{ CA1, ERC, PRC, PHC, CA3/DG, Sub }
HDDM (Ours)	{ PRC, PHC, ERC, CA3/DG, CA1, Sub }

performance of our model indicates a better ability to deal with real-world problems.

2) fMRI Hippocampus Data: We then applied our method to another real-world dataset called functional magnetic resonance imaging (fMRI) hippocampus dataset, which consisted of the resting-state signals from six brain regions of an individual [46]. It was collected from the same individual for 84 successive days and the six brain regions are perirhinal cortex (PRC), parahippocampal cortex (PHC), entorhinal cortex (ERC), subiculum (Sub), CA1, and CA3/dentate gyrus (DG). Each of the brain regions had both left and right sides, and thus, we regarded every region as a 2-D variable and we focused on investigating the causal ordering between such six brain regions. The sample size for each day is 518.

We used the anatomical connectivity between regions as a reference [47], [48]. The experimental settings and the comparisons are the same as those applied in the synthetic data. Since the data were collected for totally 84 days, we performed experiments for 84 trials and selected the variable with the maximum number of discoveries to determine the final causal order. The resulting causal orderings are shown in Table II. From Table II, overall, we see that our method and GDL-nlcorr estimate more consistent causal orderings, compared with other methods, including the first two exogenous variables, PRC and PHC. This finding also coincides with the current research, i.e., the relations between {PRC,PHC} and ERC are usually related to episodic memories, which suggest distinct roles in memory formation and retrieval [49]. Thus, it is verified that our method did obtain a better performance and have the potential to deal with real-world applications.

VI. CONCLUSION

In this article, we presented an integrated algorithm for nonlinear causal discovery, which can estimate a causal ordering in multiple high-dimensional data. We established the HDDM and derived two candidates' selection rules to reduce the identification errors. Furthermore, theoretical justification was provided to guarantee the consistency and correctness of our proposed method. Experiments on synthetic data and real-world applications demonstrate the effectiveness and superiority of our method.

REFERENCES

[1] J. Pearl, *Causality: Models, Reasoning, and Inference*. New York, NY, USA: Cambridge Univ. Press, 2000.

[2] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 2000.

[3] P. Spirtes and K. Zhang, "Causal discovery and inference: Concepts and recent methodological advances," *Appl. Informat.*, vol. 3, no. 1, 2016, Art. no. 3.

[4] P. Jonas, J. Dominik, and S. Bernhard, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[5] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Soc. Sci. Comput. Rev.*, vol. 9, no. 1, pp. 62–72, 1991.

[6] J. Pearl and T. S. Verma, "A theory of inferred causation," *Stud. Logic Found. Math.*, vol. 134, pp. 789–811, Jan. 1995.

[7] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, Nov. 2002.

[8] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, pp. 2003–2030, Dec. 2006.

[9] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 689–696.

[10] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proc. 25th Conf. Uncertainty Artif. Intell. (UAI)*, 2009, pp. 647–655.

[11] D. Janzing, P. O. Hoyer, and B. Schölkopf, "Telling cause from effect based on high-dimensional observations," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 479–486.

[12] J. Zscheischler, D. Janzing, and K. Zhang, "Testing whether linear equations are causal: A free probability theory approach," in *Proc. 27th Conf. Uncertainty Artif. Intell. (UAI)*, 2011, pp. 839–848.

[13] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic Markov condition," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5168–5194, Oct. 2010.

[14] D. Janzing *et al.*, "Information-geometric approach to inferring causal directions," *Artif. Intell.*, vol. 182, pp. 1–31, May 2012.

[15] P. Daniušis *et al.*, "Inferring deterministic causal relations," in *Proc. 26th Conf. Uncertainty Artif. Intell. (UAI)*, 2010, pp. 143–150.

[16] Z. Chen, K. Zhang, and L. Chan, "Nonlinear causal discovery for high dimensional data: A kernelized trace method," in *Proc. IEEE 13th Int. Conf. Data Mining (ICDM)*, Dec. 2013, pp. 1003–1008.

[17] R. Cai, Z. Zhang, Z. Hao, and M. Winslett, "Understanding social causalities behind human action sequences," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1801–1813, Aug. 2016.

[18] D. Entner and P. O. Hoyer, "Estimating a causal order among groups of variables in linear models," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2012, pp. 84–91.

[19] Z. Chen, K. Zhang, L. Chan, and B. Schölkopf, "Causal discovery via reproducing kernel Hilbert space embeddings," *Neural Comput.*, vol. 26, no. 7, pp. 1484–1517, 2014.

[20] F. Liu and L.-W. Chan, "Causal inference on multidimensional data using free probability theory," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 3188–3198, Jul. 2018.

[21] K. Zhang and L. Chan, "Minimal nonlinear distortion principle for nonlinear independent component analysis," *J. Mach. Learn. Res.*, vol. 9, pp. 2455–2487, Nov. 2008.

[22] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*, 2012, pp. 1255–1262.

[23] K. Zhang, B. Schölkopf, P. Spirtes, and C. Glymour, "Learning causality and causality-related learning: Some recent progress," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 26–29, Nov. 2017.

[24] N. Tagasovska, V. Chavez-Demoulin, and T. Vatter, "Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9311–9323.

[25] P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf, "Cause-effect inference by comparing regression errors," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 900–909.

[26] N. Shajarisales, D. Janzing, B. Schoelkopf, and M. Besserve, "Telling cause from effect in deterministic linear dynamical systems," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 285–294.

[27] R. Cai, J. Ye, J. Qiao, H. Fu, and Z. Hao, "FOM: Fourth-order moment based causal direction identification on the heteroscedastic data," *Neural Netw.*, vol. 124, pp. 193–201, Apr. 2020.

[28] A. Marx and J. Vreeken, "Identifiability of cause and effect using regularized regression," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 852–861.
[29] M. Kocaoglu, A. Dimakis, S. Vishwanath, and B. Hassibi, "Entropic causal inference," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–7.
[30] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: Methods and benchmarks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1103–1204, 2016.
[31] D. Janzing and B. Schölkopf, "Detecting non-causal artifacts in multivariate linear regression models," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2245–2253.
[32] F. Liu and L. Chan, "Confounder detection in high-dimensional linear models using first moments of spectral measures," *Neural Comput.*, vol. 30, no. 8, pp. 2284–2318, Aug. 2018.
[33] D. Janzing, "Causal regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 12704–12714.
[34] A. Marx and J. Vreeken, "Causal inference on multivariate and mixed-type data," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2018, pp. 655–671.
[35] S. Hu, Z. Chen, and L. Chan, "A kernel embedding–based approach for nonstationary causal model inference," *Neural Comput.*, vol. 30, no. 5, pp. 1394–1425, May 2018.
[36] K. Budhathoki and J. Vreeken, "Origo: Causal inference by compression," *Knowl. Inf. Syst.*, vol. 56, no. 2, pp. 285–307, Aug. 2018.
[37] D. Kaltenpoth and J. Vreeken, "We are not your real parents: Telling causal from confounded using MDL," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 199–207.
[38] S. Mukherjee, H. Asnani, and S. Kannan, "CCMI: Classifier based conditional mutual information estimation," in *Proc. Uncertainty Artif. Intell.*, 2020, pp. 1083–1093.
[39] D. Voiculescu, "Free probability theory: Random matrices and von Neumann algebras," in *Proc. Int. Congr. Math.* Basel, Switzerland: Springer, 1995, pp. 227–242.
[40] R. Speicher, "Free probability theory and non-crossing partitions," *Séminaire Lotharingien Combinatoire*, vol. 39, p. 38, Apr. 1997.
[41] A. Bazin, M. Couceiro, M.-D. Devignes, and A. Napoli, "Steps towards causal formal concept analysis," HAL, Lyon, France, Tech. Rep., 2021.
[42] S. Shimizu *et al.*, "DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model," *J. Mach. Learn. Res.*, vol. 12, pp. 1225–1248, Apr. 2011.
[43] A. Hyvärinen and S. M. Smith, "Pairwise likelihood ratios for estimation of non-Gaussian structural equation models," *J. Mach. Learn. Res.*, vol. 14, pp. 111–152, Jan. 2013.
[44] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[45] D. Hernandez-Lobato, P. Morales-Mombiela, D. Lopez-Paz, and A. Suarez, "Non-linear causal inference using Gaussianity measures," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 939–977, 2016.
[46] R. A. Poldrack *et al.*, "Long-term neural and physiological phenotyping of a single human," *Nature Commun.*, vol. 6, no. 1, pp. 1–15, Dec. 2015.
[47] C. M. Bird and N. Burgess, "The hippocampus and memory: Insights from spatial processing," *Nature Rev. Neurosci.*, vol. 9, no. 3, pp. 182–194, Mar. 2008.
[48] A. Ghassami, N. Kiyavash, B. Huang, and K. Zhang, "Multi-domain causal structure learning in linear systems," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 6266–6276.
[49] T. Kuhn *et al.*, "Temporal lobe epilepsy affects spatial organization of entorhinal cortex connectivity," *Epilepsy Behav.*, vol. 88, pp. 87–95, Nov. 2018.

Yan Zeng received the B.S. degree from the School of Applied Mathematics, Guangdong University of Technology, Guangzhou, China, in 2016, and the Ph.D. degree from the School of Computer, Guangdong University of Technology in 2021.

She was an Intern at the Causal Inference Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, from 2019 to 2020. She is currently a Post-Doctoral Researcher with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Her current research interests include causal discovery, reinforcement learning, matching learning, and latent variable modeling.

Zhifeng Hao (Member, IEEE) received the B.S. degree in mathematics from Sun Yat-sen University, Guangzhou, China, in 1990, and the Ph.D. degree in mathematics from Nanjing University, Nanjing, China, in 1995.

He is currently a Professor with the College of Science, Shantou University, Shantou, China. His research interests involve various aspects of algebra, machine learning, data mining, and evolutionary algorithms.

Ruichu Cai (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in computer science from South China University of Technology, Guangzhou, China, in 2005 and 2010, respectively.

He was a Visiting Student at the National University of Singapore, Singapore, from 2007 to 2009, and a Research Fellow at the Advanced Digital Sciences Center, Illinois at Singapore Pte., Singapore, from 2013 to 2014. He is currently a Professor with the School of Computer, Guangdong University of Technology, Guangzhou, and Pazhou Lab, Guangzhou. His research interests cover a variety of different topics, including causality and machine learning and their applications.

Feng Xie received the Ph.D. degree from the School of Computer Science, Guangdong University of Technology, Guangzhou, China, in 2020.

He is currently a Post-Doctoral Researcher with the School of Mathematical Sciences, Peking University, Beijing, China. His current research interests lie in causal discovery, especially in the latent causal model and its applications.

Libo Huang (Student Member, IEEE) received the B.S. degree in mathematics from Jiangxi Normal University, Nanchang, China, in 2016, and the Ph.D. degree from the School of Information Engineering, Guangdong University of Technology, Guangzhou, China, in 2021.

He was a Visiting Ph.D. Student at the Department of Electronics and Computer Engineering, Brunel University London, London, U.K., in 2019. He is currently a Post-Doctoral Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, optimization theory, signal processing, and machine learning.

Shohei Shimizu received the Ph.D. degree in statistical science engineering from Osaka University, Suita, Japan, in 2006.

He is currently a Professor with the Faculty of Data Science, Shiga University, Hikone, Japan, and the Leader of the Causal Inference Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. His current research interests include statistical methodologies for learning data generating processes, such as structural equation modeling and independent component analysis, and their application to causal inference.