Information Sciences 654 (2024) 119789

Contents lists available at ScienceDirect

Information Sciences

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

Nonlinear learning method for local causal structures

Xingyu Wu [a,b], Yan Zhong [c], Zhaolong Ling [d], Jie Yang [e], Li Li [f], Weiguo Sheng [a,*], Bingbing Jiang [a,*]

[a] School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China
[b] Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR 999077, China
[c] School of Mathematical Sciences, Peking University, Beijing 100091, China
[d] School of Computer Science and Technology, Anhui University, Hefei 230601, China
[e] Australian AI Institute, University of Technology Sydney, Sydney 2007, Australia
[f] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541214, China

ARTICLE INFO

Keywords:
Causal learning
Causal discovery
Local causal structure
Bayesian network
Markov boundary
Markov blanket

ABSTRACT

Recent years have witnessed the proliferation of causal learning techniques, aimed at extracting the abundant causal relationships embedded within observational data. In many scenarios, our primary focus lies in predicting a single target variable. In such cases, it becomes both inefficient and unnecessary to learn an entire causal network through advanced global learning methods. To address this challenge, the concept of local causal learning has been introduced to identify the direct causes and effects of a target variable of interest. However, current algorithms exhibit limitations stemming from their reliance on conditional independence tests, which only consider the linear and pairwise relationships but ignore the ubiquitous nonlinear and multivariate causality, making them lose efficacy in practical scenarios. This paper takes significant strides toward facilitating the real-world applications for local causal learning. To identify the nonlinear relations, this paper discovers the Markov boundary (MB) through calculating the minimal conditional covariance operator in reproducing kernel Hilbert space. This approach establishes a theoretical equivalence between the solution and MB. Subsequently, a nonlinear scoring mechanism for structure learning is employed based on the selected subset, yielding the optimal local causal structure. A series of extensive experiments serves to underscore the superiority of the proposed method.

1. Introduction

Causal learning aims to reveal the causal relationships among a random variable set, which could uncover the underlying generative mechanism behind the data to facilitate the interpretability and generalization ability of learning models [1]. A considerable body of research has been dedicated to predicting causality from observational data, focusing primarily on two tasks: global causal learning [2] and local causal learning [3]. As depicted in Fig. 1, global learning methods strive to acquire the complete causal graph, thereby identifying all causal relationships within a variable set and constituting the majority of research in this field. In contrast, local learning algorithms concentrate solely on the causal structure surrounding a specific target variable, aiming to uncover its direct

* Corresponding authors.
E-mail addresses: xingy.wu@polyu.edu.hk (X. Wu), zhongyan@stu.pku.edu.cn (Y. Zhong), zlling@ahu.edu.cn (Z. Ling), jie.yang-1@uts.edu.au (J. Yang), lili@guet.edu.cn (L. Li), w.sheng@ieee.org (W. Sheng), jiangbb@hznu.edu.cn (B. Jiang).

https://doi.org/10.1016/j.ins.2023.119789
Received 9 July 2023; Received in revised form 4 September 2023; Accepted 15 October 2023

Available online 20 October 2023

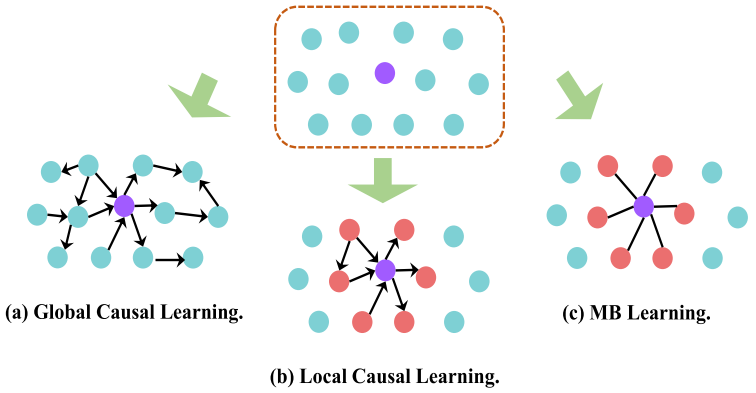

Fig. 1. The concept illustration of global causal learning, local causal learning, and MB learning. Purple circles denote the target, blue circles denote common variables, and red circles denote causal variables.

causes and effects. Local causal structures inherently retain the capacity to retain pivotal information about the target variable, which can not only provide local causality of target to promote the learning system to understand the underlying causal mechanism, but also be used as a data preprocessing step to control the learning cost and improve the interpretability. Intuitively, a global learning algorithm can also be used to obtain a local structure by reading out from the learned global network, while it is wasteful and unnecessary to solve the so large-scale NP-hard problem if we are only interested in the prediction of a certain target, especially in real-world large-scale network [4].

Several local causal learning methodologies have been developed to identify the local structure surrounding a given target variable without the need to learn the entire causal network. Typically, these algorithms begin by identifying the Markov boundary (MB) set[1] or MB subset of the target. Subsequently, they sequentially uncover the MB (sub)sets of variables connected to the target while simultaneously framing the directed paths between the target and these selected variables, ultimately constructing the local causal structure [6]. Compared to global learning methods, these algorithms substantially reduce time complexity by focusing on the MB set instead of the entire variable set [4]. However, a weakness inherent in the MB discovery procedure is its inability to consider multivariate nonlinear causality, despite its prevalence in real-world data. Current MB discovery techniques typically assess whether a variable should be included in the MB set one by one [7], using a conditional independence test between pairs of variables. This approach fails to capture multivariate nonlinear relationships (e.g., the logical operation 'XOR') and can only identify pairwise causality. Furthermore, the conditional independence test's performance is constrained by the size of the conditioning set, resulting in reduced accuracy, especially with limited training instances.

A feasible solution for the identification of multivariate causality is to directly implement the test between variable combinations and the target, while various combinations will bring prohibitively expensive calculation, and the conditional independence test between target and variable subset also declines the reliability of identification since the exponentially growing value space reduces the sample size per degree of freedom for hypothesis testing in conditional independence test [8]. Analogously, another solution, employing a score function for structure learning that considers nonlinear relations, also suffers from the time-consuming process due to the searching space on the entire variable set, whose time complexity is comparable to the global learning approaches.

In light of these challenges, we propose a novel approach for kernel-based local causal learning, departing from the conventional conditional independence test. In this paper, we map the variable space and target space to the reproducing kernel Hilbert space (RKHS) [9] to consider both pairwise and multivariate relationships. The statistical concept, conditional covariance operator [10], is introduced to describe the conditional dependence and independence between variables. We will theoretically establish that the MB of a target is equivalent to the variable subset that minimizes the conditional covariance operator, enabling our proposed methods to learn the MB variables by solving an optimization problem in RKHS. Based on the skeleton of the local structure constructed by the learned MB, a nonlinear Bayesian network score function is employed to determine the orientation of the optimal local causal structure, where the calculation complexity is significantly decreased as the MB scale is much smaller than the entire variable set. The main contributions in this paper are summarized as follows:

- We theoretically demonstrate the equivalence between the MB and the minimal conditional covariance operator, paving the way for a novel kernel-based MB discovery strategy that minimizes the conditional covariance operator in RKHS.
- A novel local causal learning method is proposed based on the kernel MB discovery strategy, which could retrieve the local causal structure around the target, and simultaneously identify both the linear and nonlinear (including multivariate) causality. To the best of our knowledge, it is the first nonlinear local causal learning algorithm.
- Extensive experiments on synthetic and real-world data sets validate the practicability and superiority of the proposed approaches in the MB discovery and local causal learning tasks.

[1] In a causal network, the MB set of a target comprises its direct causes (parent nodes), direct effects (child nodes), and other direct causes of its direct effects (spouse nodes) [5].

2. Related work

Numerous extensive studies [1] have consistently highlighted the superiority of causality over correlation across diverse learning scenarios. Causal learning, alternatively referred to as causal discovery [11], has garnered significant attention as a means to reveal causal relationships within a set of variables, to facilitate the interpretability and generalization ability of learning models [11]. A substantial body of research has been dedicated to the task of predicting causality from observational data [11], broadly categorized into two sub-areas: global causal discovery [2] and local causal discovery [3]. In the following sections, we delve into related work in these two areas, with additional discussion on Markov Boundary (MB) discovery, a fundamental technique in local causal learning algorithms.

2.1. Global causal learning methods

Global causal discovery tries to reveal all the causal relationships among a variable set simultaneously, that is, these algorithms construct the entire causal Bayesian network (directed acyclic graph, DAG) for all variables, where a directed edge denotes the causality from a cause variable to an effect variable. Some pioneer approaches are proposed using interventions or randomized experiments [12], whereas only passive observation can be performed but active intervention cannot be implemented in most cases due to the limitations of experimental technology. This traditional way is replaced with causal learning from purely observational data [13], which can be roughly divided into four types: constraint-based approaches, score-based approaches, functional causal models-based approaches, and gradient-based approaches.

Early attention in this field was directed towards constraint-based methods, which predominantly exploit conditional independence relationships within the data to uncover underlying causal structures. These algorithms, while not providing complete causal information, yield sets of causal structures that adhere to the same conditional independence criteria, known as Markov equivalence classes. Score-based approaches have emerged to select the most appropriate structure from these Markov equivalence classes [14], with examples like the greedy equivalence search (GES) algorithm [15]. However, finding the causal graph with the highest score is both NP-hard and NP-complete, resulting in a local optimum in most instances. To select the unique optimal structure, functional causal models [16] implement causal discovery by constructing a structural equation between cause and effect to represent the causal order, which has demonstrated their superiority when searching the optimal structures among the Markov equivalent structures, e.g., the Linear Non-Gaussian Acyclic Model (LiNGAM) [16]. Recent innovations have introduced gradient-based approaches for global learning [17], transforming the combinatorial optimization problem into a continuous optimization problem solvable through gradient descent.

2.2. Local causal learning methods

In certain practical applications, the focus narrows to the causal structure around a specific target variable, obviating the need for the time-consuming global learning procedure that encompasses all causal relationships. To tackle the local causal discovery, some algorithms are proposed to distinguish the direct causes and direct effects of a target variable [3]. While a significant amount of research has been dedicated to global learning algorithms, there has been a limited number of proposed algorithms specifically designed for local causal structure learning.

The pioneering algorithm developed for local causal structure learning is PCD-by-PCD (where PCD stands for Parents, Children, and Descendants) [4]. It first identifies the PCD set of a target variable and subsequently uncovers the PCD sets of variables directly linked to the target. The algorithm records V-structures to assist in determining the orientation of edges involving the target variable until all its direct causes and effects are identified. Another algorithm, Causal Markov Blanket (CMB) [3], initially learns an MB set for the target variable and subsequently orients edges by monitoring changes in conditional dependence and independence during the MB discovery process. CMB sequentially learns the MB sets of variables connected to the target and constructs local structures along the paths starting from the target variable until the direct causes and effects of the target are identified. The Efficient Local Causal Structure (ELCS) learning algorithm [18] introduces a new concept called N-structures to enhance MB discovery, thereby improving the time efficiency of local learning. However, none of these algorithms consider the nonlinear relationships present in the data. This paper proposes a novel local causal learning algorithm that combines linear and nonlinear causality based on kernel MB discovery to uncover local structures. This algorithm addresses the MB discovery as a subtask of local causal discovery, and we delve into the related work of MB discovery in the following subsections.

2.3. Markov boundary discovery

The Markov boundary (MB) of a target comprises its direct causes (parents), direct effects (children), and other direct causes of its direct effects within the causal Bayesian network [19]. Thus, MB discovery represents a subprocedure of causal discovery [19,20], yielding a skeleton that describes variable relationships without orientation around a specific target. MB provides a complete picture of the local causal structure around the target variable, which possesses a superior property: For target T and its MB $\boldsymbol{MB} \subset \boldsymbol{X}$, all other variables $X \in \boldsymbol{X} - \boldsymbol{MB}$ are independent of T conditioned on $\boldsymbol{MB}$, and any subsets of $\boldsymbol{MB}$ do not satisfy the condition. This property underscores that the MB of a target encompasses all predictive causal information about the target and has been applied in feature selection techniques, known as “causal feature selection” as proposed by Guyon et al. [21].

These constraint-based methods have found application across various scenarios, including multi-label data [22], biomedical data [8], and streaming data [23]. However, the mainstream techniques for conditional independence tests are constrained by the size of the conditioning set and the number of instances, rendering constraint-based methods unsuitable for scenarios featuring large MB sizes and limited samples. In contrast, score-based algorithms [24,25] employ scoring functions [26] and a greedy search strategy to evaluate the compatibility between the probability distribution in training data and the learned causal graph. These algorithms, often characterized by high time complexity, are less suitable for large-scale variable sets.

Existing MB learning algorithms [6] are roughly divided into constraint-based and score-based algorithms. Constraint-based methods [27,28] account for the majority of MB research, as the focus of this area, which learn MB via mining the conditional dependence and independence in the variable set. These constraint-based methods have found application across various scenarios, including multi-label data [22], biomedical data [8], and streaming data [23]. However, the mainstream techniques for conditional independence tests are constrained by the size of the conditioning set and the number of instances, rendering constraint-based methods unsuitable for scenarios featuring large MB sizes and limited samples. To overcome this flaw, score-based algorithms [24,25] adopt a scoring function [26] and a greedy search method to measure the fitness between the probability distribution in training data and learned causal graph. These algorithms are usually with high time complexity and are not suitable for the large-scale variable set.

3. Local causal learning in RKHS

In this section, the capital letters (such as X) represent random variables, and the capital bold italic letters (such as $\boldsymbol{Z}$) denote variable sets. Specifically, $\boldsymbol{U} = \{X_1, X_2, \cdots, X_n\}$ denotes the entire variable set, $T \in \boldsymbol{U}$ denotes the target, and $\boldsymbol{s}_j = \{x_1, x_2, \cdots, x_n\}(j = 1, \cdots, m)$ denotes an instance. In addition, the symbol $X \not\perp Y|\boldsymbol{Z}$ ($X \perp Y|\boldsymbol{Z}$) represents that variables X and Y are conditionally (in)dependent given a variable set $\boldsymbol{Z}$.

As mentioned in Section 1, the foundation of local causal learning lies in MB discovery. Therefore, to identify nonlinear multivariate relationships in the data, we need to equip MB discovery techniques with this capability. Kernel methods [9] have demonstrated their effectiveness in addressing nonlinear problems across various scenarios. In this paper, we leverage kernel-based techniques to uncover nonlinear causality. Typically, the MB variables are represented in the original Euclidean space. Consequently, we can not directly obtain the MB set via searching in the mapped RKHS. In the following discussion, we take a detour to surmount this obstacle using a statistical concept within the RKHS, namely the conditional covariance operator [10]. We first provide the equivalence analyses between MB and the conditional covariance operator in the RKHS.

The covariance operator was first used in Banach spaces [9] and introduced to RKHS [10] later, as an extension based on the covariance in Euclidean space. We use the symbol $\mathcal{H}_{\mathcal{X}}$ to represent the RKHS associated with the Euclidean space $\mathcal{X}$. The cross-covariance operator for the random variable pair (X, Y) is the mapping from $\mathcal{H}_{\mathcal{X}}$ to $\mathcal{H}_{\mathcal{Y}}$. For each $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$:

$$\begin{aligned}\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_Y} =& Cov[f(X), g(Y)] \\ =& E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)],\end{aligned} \tag{1}$$

where Σ_{YX} denotes the cross-covariance operator, which could be understood as the covariance matrix over the feature maps $\Phi_{\mathcal{Y}}(Y)$ and $\Phi_{\mathcal{X}}(X)$. Under the concept of cross-covariance operator, the dependence relationships between two variables could be transformed to the zero cross-covariance operator, that is, $Cov[f(X), g(Y)] = 0$:

$$X \perp Y \iff \Sigma_{YX} = 0 \iff Cov[f(X), g(Y)] = 0. \tag{2}$$

Conditional dependence can express more essential relations among variables, including the causality. Based on the cross-covariance operator, we provide the definition of conditional covariance operator [9] as follows to establish the formulation of conditional independence in RKHS. Since $\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$, then we can calculate the conditional covariance operator $\Sigma_{YY|X}$ on $\mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{Y}}$ using:

$$\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \tag{3}$$

In the square-integrable space $L^2(P_X)$ on X, the following equation is true for $\forall g \in \mathcal{H}_Y$ if the sum $\mathcal{H}_{\mathcal{Y}} + \mathbb{R}$ is dense according to [10]:

$$\forall g \in \mathcal{H}_{\mathcal{Y}}, \left\langle g, \Sigma_{YY|X} g\right\rangle_{\mathcal{H}_{\mathcal{Y}}} = E_X[D_{Y|X}[g(Y)|X]] \tag{4}$$

Additionally, the residual error of $g \in \mathcal{H}_Y$ can be characterized by the conditional covariance operator as

$$\left\langle g, \Sigma_{YY|X} g\right\rangle_{\mathcal{H}_Y} = \inf_{f \in \mathcal{H}_X} E_{XY}[(g(Y) - E_Y[g(Y)]) - (f(X) - E_X[f(X)])]^2. \tag{5}$$

With this mathematical foundation in place, we can establish a connection between the Markov Boundary (MB) and the conditional covariance operator within the Reproducing Kernel Hilbert Space (RKHS). For a variable set $\boldsymbol{U}$ and a target variable $T \in \boldsymbol{U}$, the MB of T must satisfy the independence property ($\boldsymbol{U} - \boldsymbol{Z} \perp T|\boldsymbol{Z}$) and the minimality requirement (no subsets of $\boldsymbol{Z}$ should satisfy the independence property). We will demonstrate that the minimal conditional covariance operator conforms to these two conditions, respectively. In this context, we employ the kernel trick to assess the causal information of a variable within the RKHS. The original sample space $\varsigma = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_m\}$ and target space $\tau = \{t_1, t_2, \cdots, t_m\}$ are mapped into the RKHS $\mathcal{H}_\varsigma$ and $\mathcal{H}_\tau$, respectively, with two

measurable positive definite kernels $k_\varsigma : \varsigma \times \varsigma \to \mathbb{R}$ and $k_\tau : \tau \times \tau \to \mathbb{R}$, which are both chosen as radial basis function in this paper as follows:

$$k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}. \tag{6}$$

And the two corresponding kernel matrixes are represented as $(\boldsymbol{K}_\varsigma)_{ij} = k_\varsigma(\boldsymbol{s}_i, \boldsymbol{s}_j)$ and $(\boldsymbol{K}_\tau)_{ij} = k_\tau(t_i, t_j)$.

According to Eq. (4), for $\forall g \in \mathcal{H}_\mathcal{T}$ and $\boldsymbol{Z} \subset \boldsymbol{U}$:

$$\left\langle g, \Sigma_{TT|\boldsymbol{Z}} g \right\rangle_{\mathcal{H}_\tau} = E_{\boldsymbol{Z}}[D_{T|\boldsymbol{Z}}[g(T)|\boldsymbol{Z}]] \tag{7}$$

$$\left\langle g, \Sigma_{TT|\boldsymbol{U}} g \right\rangle_{\mathcal{H}_\tau} = E_{\boldsymbol{U}}[D_{T|\boldsymbol{U}}[g(T)|\boldsymbol{U}]] \tag{8}$$

The conditional covariance operators on $\boldsymbol{Z} \subset \boldsymbol{U}$ and $\boldsymbol{U}$ can be compared according to Eq. (5). Because the infimum within the scope of a subset $\boldsymbol{Z}$ should exceed the infimum within the scope of the entire variable set $\boldsymbol{U}$, we have

$$\left\langle g, \Sigma_{TT|\boldsymbol{U}} g \right\rangle_{\mathcal{H}_\tau} \leq \left\langle g, \Sigma_{TT|\boldsymbol{Z}} g \right\rangle_{\mathcal{H}_\tau} \tag{9}$$

This inequality implies that $\Sigma_{TT|\boldsymbol{U}} \preceq \Sigma_{TT|\boldsymbol{Z}}$. Considering the condition for equality in Eq. (9), we analyze the statistical properties of the MB when the equality holds, i.e., $\Sigma_{TT|\boldsymbol{X}} = \Sigma_{TT|\boldsymbol{Z}}$. The variance $D_{T|\boldsymbol{Z}}[g(T)|\boldsymbol{Z}]$ in Eq. (7) could be decomposed according to the law of total variance:

$$D_{T|\boldsymbol{Z}}[g(T)|\boldsymbol{Z}] = E_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[D_{T|\boldsymbol{U}}[g(T) \mid (\boldsymbol{U}-\boldsymbol{Z}) \cup \boldsymbol{Z}]\right] + D_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[E_{T|\boldsymbol{U}}[g(T) \mid (\boldsymbol{U}-\boldsymbol{Z}) \cup \boldsymbol{Z}]\right] \tag{10}$$

Calculating the expectation value of both sides in Eq. (10), we have:

$$E_{\boldsymbol{Z}}\left[D_{T|\boldsymbol{Z}}[g(T)|\boldsymbol{Z}]\right] = E_{\boldsymbol{U}}\left[D_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]\right] + E_{\boldsymbol{Z}}\left[D_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]\right]\right] \tag{11}$$

According to Eq. (7) and Eq. (8), Eq. (11) can be simplified as by substituting the two equations:

$$\left\langle g, \Sigma_{TT|\boldsymbol{Z}} g \right\rangle_{\mathcal{H}_\tau} = \left\langle g, \Sigma_{TT|\boldsymbol{U}} g \right\rangle_{\mathcal{H}_\tau} + E_{\boldsymbol{Z}}\left[D_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]\right]\right] \tag{12}$$

where $E_{\boldsymbol{Z}}\left[D_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]\right]\right] = 0$ in the case of $\left\langle g, (\Sigma_{TT|\boldsymbol{Z}} - \Sigma_{TT|\boldsymbol{U}}) g \right\rangle_{\mathcal{H}_\tau} = 0$. Since the variance is nonnegative, $E_{\boldsymbol{Z}}\left[D_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]\right]\right] = 0$ is equivalent to $D_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]\right] = 0$ and $E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]$ is a constant. Therefore, $E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]$ is independent of the variable set $\boldsymbol{U}$ conditioned on $\boldsymbol{Z}$ in the case of $\Sigma_{TT|\boldsymbol{Z}} = \Sigma_{TT|\boldsymbol{U}}$, that is:

$$E_{\boldsymbol{Z}}\left[D_{(\boldsymbol{U}-\boldsymbol{Z})|\boldsymbol{Z}}\left[E_{T|\boldsymbol{U}}[g(T) \mid \boldsymbol{U}]\right]\right] = 0 \iff T \perp \boldsymbol{U}|\boldsymbol{Z} \tag{13}$$

which indicates that the variable subset $\boldsymbol{Z}$ satisfying the $\Sigma_{TT|\boldsymbol{Z}} = \Sigma_{TT|\boldsymbol{U}}$ is the MB of target T. According to Eq. (9), the objective of MB discovery can be taken as the minimization of the $\Sigma_{TT|\boldsymbol{Z}}$. Formally,

$$MB(T) = \arg\min_{\boldsymbol{Z} \subset \boldsymbol{X}} \Sigma_{TT|\boldsymbol{Z}} \tag{14}$$

In this paper, we provide two methods to estimate the value of $\Sigma_{TT|\boldsymbol{Z}}$.

(M1) **Estimation through determinant of** $\Sigma_{TT|\boldsymbol{Z}}$: In an ordered set of positive definite matrices, the value of $\Sigma_{TT|\boldsymbol{Z}}$ could be estimated through its determinant. According to the Schur theorem [29],

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} E & 0 \\ -C^{-1}B^T & E \end{pmatrix} = \begin{pmatrix} A - BC^{-1}B^T & B \\ 0 & C \end{pmatrix} \tag{15}$$

Take the determinant for both sides of Eq. (15), then:

$$\det\left(A - BC^{-1}B^T\right) = \det\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} / \det C \tag{16}$$

Substitute $A = \hat{\Sigma}_{TT}$, $B = \hat{\Sigma}_{T\boldsymbol{Z}}$, and $C = \hat{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}}$ into Eq. (16), then

$$\det \hat{\Sigma}_{TT|\boldsymbol{Z}} = \frac{\det \hat{\Sigma}_{[T\boldsymbol{Z}][T\boldsymbol{Z}]}}{\det \hat{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}}} \tag{17}$$

where the value of $\hat{\Sigma}_{TT}$, $\hat{\Sigma}_{T\boldsymbol{Z}}$, and $\hat{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}}$ could be easily obtained with a given kernel matrix according to [10] like:

$$\hat{\Sigma}_{T\boldsymbol{Z}} = \frac{1}{n} G_T G_{\boldsymbol{Z}} \tag{18}$$

in which $G_{\boldsymbol{Z}}$ is calculated as follows:

$$G_{\boldsymbol{Z}} = (I_n - \frac{1}{n} 1_n 1_n^T) \boldsymbol{K}_{\boldsymbol{Z}} (I_n - \frac{1}{n} 1_n 1_n^T) \tag{19}$$

where $\boldsymbol{K}_{\boldsymbol{Z}}$ is the kernel matrix with a lower-dimensional variable set $\boldsymbol{Z} \subset \boldsymbol{U}$, I_n and 1_n represent the $n \times n$ identity matrix and $n \times 1$ vector with all ones, respectively. And G_T is calculated as follows:

Algorithm 1 The LC-KMB Algorithm.

1: **Input:** Target T, variable set $\boldsymbol{U}$, training instances $\mathbb{D}$, parameters α, β.
2: Calculate $\boldsymbol{K}_\varsigma$, $\boldsymbol{K}_\tau$, and $G_{\boldsymbol{Z}}$ (Eq. (19)) with initialization of $\boldsymbol{Z} = \boldsymbol{U}$.
3: **Repeat**
4: $X \leftarrow \arg\min_{X \in \boldsymbol{Z}} \Sigma_{TT|\boldsymbol{Z}-\{X\}}$
5: $\boldsymbol{Z} \leftarrow \boldsymbol{Z} - \{X\}$.
6: Update $\boldsymbol{K}_\varsigma$, $\boldsymbol{K}_\tau$ and $G_{\boldsymbol{Z}}$ according to Eq. (19).
7: **Until** $|\boldsymbol{Z}| \le \alpha|\boldsymbol{U}|$
8: **Repeat**
9: For each $X_i \in \boldsymbol{Z}$, determine the λ_i with Armijo rule, and calculate $\omega_i := \omega_i - \lambda_i \frac{\partial \Theta}{\partial \omega_i}$ with Eq. (17) or (21)
10: **Until** Convergence of the objective function.
11: $\boldsymbol{Z} = \{X_i | sigmoid(\omega_i) > \beta\}$.
12: $\mathbb{G} \leftarrow \arg\max_{\mathbb{G}} \text{ScoreDAG}(\mathbb{D}, \mathbb{G}, \boldsymbol{Z} \cup \{T\})$.
13: **Output:** Direct causes and effects in $\mathbb{G}$.

$$(G_T)_{ij} = (\boldsymbol{K}_\tau)_{ij} + \frac{1}{m}\sum_{a=1}^{m}(\boldsymbol{K}_\tau)_{aj} - \frac{1}{m}\sum_{b=1}^{m}(\boldsymbol{K}_\tau)_{ib} + \frac{1}{m^2}\sum_{a=1}^{m}\sum_{b=1}^{m}(\boldsymbol{K}_\tau)_{ab}, \tag{20}$$

(M2) **Estimation through trace of** $\Sigma_{TT|\boldsymbol{Z}}$: According to Eq. (3), we formalize the trace of $\Sigma_{TT|\boldsymbol{Z}}$ as:

$$\text{Tr}(\hat{\Sigma}_{TT|\boldsymbol{Z}}) = \text{Tr}\left[\hat{\Sigma}_{TT} - \hat{\Sigma}_{T\boldsymbol{Z}}\left(\hat{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}} + \sigma I_m\right)^{-1}\hat{\Sigma}_{\boldsymbol{Z}T}\right] \tag{21}$$

where σI_m ($\sigma > 0$) is a regularization term to enable operator inversion, similar to Tikhonov regularization [30]. Similar to the calculation of determinant, we substitute $\hat{\Sigma}_{TT}$, $\hat{\Sigma}_{T\boldsymbol{Z}}$, and $\hat{\Sigma}_{\boldsymbol{Z}\boldsymbol{Z}}$ into Eq. (21), and obtain:

$$\text{Tr}(\hat{\Sigma}_{TT|\boldsymbol{Z}}) = \frac{1}{n}\text{Tr}\left[G_T - G_{\boldsymbol{Z}}\left(G_{\boldsymbol{Z}} + n\varepsilon_n I_n\right)^{-1} G_T\right] \tag{22}$$

in which the matrix can be further simplified as:

$$\begin{aligned} G_T - G_{\boldsymbol{Z}}\left(G_{\boldsymbol{Z}} + n\varepsilon_n I_n\right)^{-1} G_T =& [\left(G_{\boldsymbol{Z}} + n\varepsilon_n I_n\right)\left(G_{\boldsymbol{Z}} + n\varepsilon_n I_n\right)^{-1} - G_{\boldsymbol{Z}}\left(G_{\boldsymbol{Z}} + n\varepsilon_n I_n\right)^{-1}]G_T \\ =& n\varepsilon_n I_n\left(G_{\boldsymbol{Z}} + n\varepsilon_n I_n\right)^{-1} G_T \end{aligned} \tag{23}$$

And the estimated trace is calculated as:

$$\text{Tr}(\hat{\Sigma}_{TT|\boldsymbol{Z}}) = \varepsilon_n \text{Tr}\left[G_T\left(G_{\boldsymbol{Z}} + n\varepsilon_n I_n\right)^{-1}\right] \tag{24}$$

where the G_T and $G_{\boldsymbol{Z}}$ can be calculated using Eq. (19) and Eq. (20).

The kernel MB representation promotes us to propose a Local Causal learning algorithm based on Kernel MB discovery (LC-KMB), as shown in Algorithm 1. The core concept behind LC-KMB is to initially learn the MB within the RKHS and subsequently employ a nonlinear score function to search for the optimal local MB structure. Thus, the LC-KMB algorithm consists of three primary steps:

(1) Lines 2-7: LC-KMB first use the value of $\Sigma_{TT|\boldsymbol{Z}-\{X\}}$ to quickly remove some noncausal variables, in which the parameter α in line 7 is the percentage of the filtered variables; Line 4 involves evaluating the possibility of each variable as a potential MB variable, employing Equation $\Sigma_{TT|\boldsymbol{Z}-\{X\}}$, alongside the established monotonic partial ordering relation $\Sigma_{TT|\boldsymbol{U}} \preceq \Sigma_{TT|\boldsymbol{Z}}$ as substantiated in Eq. (9). This relationship essentially signifies that when a variable X is excluded from the variable set, a smaller value of $\Sigma_{TT|\boldsymbol{Z}-X}$ corresponds to a lower likelihood of X being the MB variable.

(2) Lines 8-11: In this part, the MB is learned by minimizing the objective function $\Sigma_{TT|\boldsymbol{Z}}$ as defined in Eq. (14). In line 9, $\Omega = \omega_1, \omega_2, \ldots, \omega_n$ signifies whether a variable is included in the discovered MB, where $\omega_i = 1$ indicates that X_i is included, and 0 otherwise. Consequently, the learned MB set is represented as $\boldsymbol{Z} = \Omega \odot \boldsymbol{U}$. The conditional covariance operator $\Sigma_{TT|\boldsymbol{Z}}$ is denoted as $\Theta(\Omega)$, which can be computed using the determinant in Eq. (17) or the trace in Eq. (21), designated as LC-KMB$_d$ and LC-KMB$_t$, respectively. To utilize gradient descent for minimizing $\Theta(\Omega)$, Ω ($\omega_i \in 0, 1$) is replaced with $sigmoid(\Omega)$ ($\omega_i \in \mathbb{R}$), and the Armijo rule is employed to adaptively determine the step sizes in each iteration. Line 11 outputs the discovered MB set, controlled by the parameter β.

(3) Lines 12: Building upon the learned MB, a nonlinear score function for Bayesian network structure learning (e.g., the score function proposed in [31]) can be utilized to construct a compact network centered around the target variable. Since the MB is learned within the mapped RKHS, it accounts for both pairwise and multivariate relationships, ensuring the inclusion of all direct causes and effects in the discovered MB. Consequently, in theory, any score function capable of identifying nonlinear relationships can uncover the correct structure. Despite these algorithms being computationally demanding, the size of the MB is significantly smaller than that of the entire variable set, resulting in a more efficient structure learning process, which differs from the considerations in global learning.

4. Experiments

In this section, we present the experimental findings derived from applying the LC-KMB algorithm to synthetic and real-world datasets. We start by demonstrating the effectiveness of the MB discovery subprocedure in LC-KMB in Section 4.1. Subsequently,

Table 1
Details of the standard benchmark Bayesian network data sets.

Data set	Domain range	#Variables	#Edges	Max In/Out Degree	Min/Max $\|PC\|$
Alarm	2-4	37	46	4/5	1/6
Alarm3	2-4	111	149	4/5	1/6
Child	2-6	20	25	2/7	1/8
Child3	2-6	60	79	3/7	1/8
Gene	3-5	801	972	4/10	0/11
Insurance	2-5	27	52	3/7	1/9
Insurance3	2-5	81	163	4/7	1/9
Pigs	3-3	441	592	2/39	1/41

Sections 4.2 and 4.3 conduct extensive experiments on synthetic datasets and real-world causal network datasets, respectively, to assess the performance of local causal structure learning compared to state-of-the-art local and global causal learning algorithms. Finally, in Section 4.4, we perform an experiment on the electroencephalography (EEG) dataset, SEED [32], to further validate the efficacy of the direct causes and effects discovered by LC-KMB.

4.1. MB discovery in standard causal network

In this subsection, we delve into the impact of MB discovery accuracy on the performance of LC-KMB, given that LC-KMB identifies nonlinear causal relationships by learning MBs in RKHS. Specifically, we demonstrate the effectiveness and superiority of LC-KMB compared to existing MB discovery algorithms.

Dataset Description: To assess the accuracy of MB discovery, we selected ten commonly-used standard Bayesian network datasets for evaluation in this subsection. These datasets are chosen because the underlying causal mechanisms are known based on their adjacent matrix, allowing for a comparison between the learned MB and the true MB to calculate the F_1 score. These networks are primarily derived from real-world decision support systems, covering various real-life applications in fields such as medicine, financial modeling, and animal breeding. Table 1 provides statistical information about these networks, reflecting differences in size, density, and data quality.

Compared Algorithms: In our evaluation, we compared six state-of-the-art MB discovery algorithms, each belonging to different categories. Four constraint-based algorithms were included: the direct method IAMB [33], and the divide-and-conquer methods HITON-MB [34], CCMB [27], and SRMB [28]. These algorithms ascertain MB by examining conditional dependence and independence within the variable set, using the G^2-test [35] for conditional independence testing in our experiments. Additionally, two score-based MB discovery algorithms, SLL [24] and S^2TMB [25], were selected.

Performance Comparison: All compared MB learning algorithms as well as LC-KMB$_d$ and LC-KMB$_t$ (only lines 1-11) are executed for each variable in these datasets with 500 instances and are repeated 10 times with different training instances. The commonly used evaluation metric F_1 scores are chosen to measure the accuracy of the discovered MB. Fig. 2 shows the accuracy performance of various MB discovery algorithms on different datasets, which demonstrates that both the LC-KMB$_d$ and LC-KMB$_t$ consistently perform better than others on all datasets. By transforming the MB discovery problem to the minimization of the conditional covariance operator, LC-KMB can take more comprehensive variable causality into consideration. The identified multivariate nonlinear causality helps LC-KMB$_d$ and LC-KMB$_t$ beat the opponent. Moreover, we also notice that LC-KMB$_t$ achieves better performance than LC-KMB$_d$ in large-scale datasets, indicating a better estimation of the conditional covariance operator by calculating its trace in this case.

4.2. Local causal learning in synthetic nonlinear data

In this subsection, we evaluate the effectiveness of LC-KMB in the local causal learning task, specifically in nonlinear data, to validate the primary contributions of this paper.

Dataset Description: For this study, we generated datasets by sampling from synthetic Bayesian networks using the simulation methodology outlined in [36]. This approach allows us to assess the performance of different methods under controlled conditions, where we have precise knowledge of the underlying mechanism and all MB variables associated with each target. To test LC-KMB's effectiveness in a nonlinear environment, we adjusted the proportion of nonlinear relationships (including multivariate relationships), denoted as p_n. The simulated Bayesian network consists of 50 variables and 1000 training samples, consistent across all experiment groups.

Compared Algorithms: We compare the proposed LC-KMB$_d$ and LC-KMB$_t$ with two state-of-the-art local learning approaches, CMB [3] and PCD-by-PCD [4], as well as two global causal structure learning algorithms, MMHC [2] and NOTEARS [17]. CMB and PCD-by-PCD are classic local learning methods based on MB discovery, using conditional independence tests to capture causality in the data. The G^2-test [35] is the chosen conditional independence test for our experiments. MMHC and NOTEARS represent traditional and state-of-the-art global learning methods, respectively. MMHC constructs a causal skeleton using a conditional independence test and employs a greedy search method alongside a Bayesian network scoring function to find the optimal structure, similar to LC-KMB. NOTEARS transforms the combinatorial optimization problem into a continuous one and solves it using gradient descent. For handling nonlinear cases, we adopted the scoring function proposed in [31].

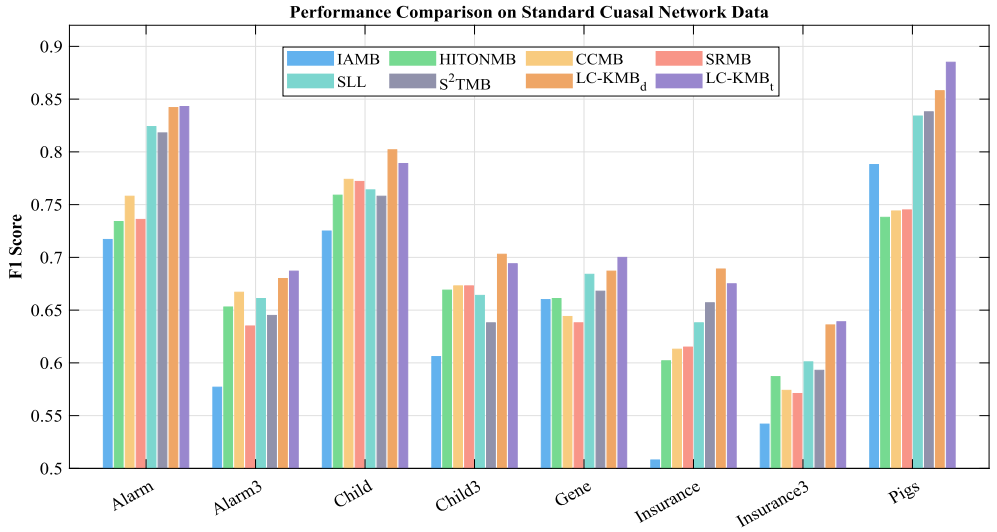

Fig. 2. The F_1-score of discovered MB on standard causal network data sets.

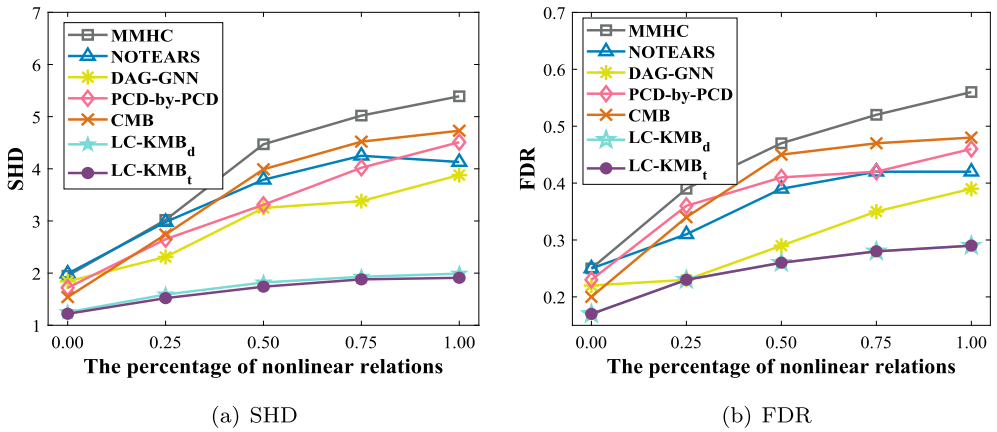

Fig. 3. The SHD and FDR of local causal structure learning achieved by LC-KMB and other state-of-the-art local and global learning algorithms on synthetic nonlinear data.

Performance Comparison: In this experiment, we assess performance using three key metrics: SHD (Structural Hamming Distance), FDR (False Discovery Rate), and $Time$ (execution time). SHD quantifies the total number of errors in the output structure, encompassing undirected edges, reversed edges, missing edges, and extra edges. A lower SHD value indicates a more accurate structure. FDR measures the number of false edges in the output divided by the total number of edges produced by the algorithm. It provides insight into the precision of the algorithm's results. $Time$ indicates the average execution time of the algorithms, measured in seconds. A shorter execution time is preferable.

Fig. 3 illustrates the performance variations of each compared algorithm as the percentage of nonlinear relationships in the datasets changes. The results clearly demonstrate the consistent superiority of LC-KMB$_d$ and LC-KMB$_t$ over the other compared algorithms. Notably, as the proportion of nonlinear relationships increases, the accuracy of existing algorithms tends to decline, although to varying degrees. In contrast, LC-KMB maintains stable performance across different experimental scenarios and consistently outperforms other MB learning methods. While LC-KMB's performance may experience a slight decline with higher proportions of nonlinear relationships, this decline is significantly smaller compared to the comparison algorithms. These observations suggest that minimizing the conditional covariance operator in MB discovery enhances LC-KMB's ability to identify nonlinear and multivariate relationships, resulting in more accurate discoveries of direct causes and effects.

4.3. Local causal learning in standard causal network

To further demonstrate the effectiveness in real-world causal learning scenarios, we implement the aforementioned local causal learning algorithms on the standard causal network data. The same experimental settings and evaluation metrics are used in this subsection.

Table 2
SHD comparison on the standard causal network datasets.

Datasets	MMHC	NOTEARS	DAG-GNN	PCD-by-PCD	CMB	LC-KMB$_d$	LC-KMB$_t$
Alarm	4.37 ± 0.21	3.19 ± 0.19	2.11 ± 0.14	1.42 ± 0.13	1.45 ± 0.14	1.09 ± 0.12	**1.05 ± 0.10**
Alarm3	5.17 ± 0.14	6.25 ± 0.11	5.19 ± 0.16	1.99 ± 0.08	1.94 ± 0.06	1.62 ± 0.11	**1.53 ± 0.08**
Child	3.47 ± 0.17	3.69 ± 0.21	2.35 ± 0.11	1.69 ± 0.19	1.42 ± 0.25	**0.97 ± 0.18**	1.04 ± 0.23
Child3	3.99 ± 0.17	3.71 ± 0.13	2.34 ± 0.11	1.75 ± 0.08	1.93 ± 0.12	1.34 ± 0.09	**1.32 ± 0.11**
Gene	-	-	-	-	-	0.81 ± 0.04	**0.78 ± 0.03**
Insurance	5.74 ± 0.25	5.11 ± 0.17	3.94 ± 0.18	2.45 ± 0.21	2.59 ± 0.20	**2.40 ± 0.23**	2.41 ± 0.19
Insurance3	4.73 ± 0.09	9.25 ± 1.25	6.99 ± 1.03	2.73 ± 0.09	3.05 ± 0.11	3.11 ± 0.05	**2.65 ± 0.06**
Pigs	6.93 ± 0.04	2.99 ± 0.06	-	-	-	0.89 ± 0.02	**0.65 ± 0.03**

Table 3
FDR comparison on the standard causal network datasets.

Datasets	MMHC	NOTEARS	DAG-GNN	PCD-by-PCD	CMB	LC-KMB$_d$	LC-KMB$_t$
Alarm	0.56 ± 0.05	0.49 ± 0.04	0.18 ± 0.04	0.21 ± 0.03	0.38 ± 0.07	0.15 ± 0.04	**0.13 ± 0.03**
Alarm3	0.62 ± 0.03	0.57 ± 0.03	0.25 ± 0.02	0.23 ± 0.03	0.41 ± 0.04	0.21 ± 0.03	**0.20 ± 0.02**
Child	0.46 ± 0.04	0.69 ± 0.06	0.28 ± 0.11	0.35 ± 0.09	0.36 ± 0.12	**0.23 ± 0.07**	0.25 ± 0.06
Child3	0.44 ± 0.02	0.53 ± 0.01	0.31 ± 0.02	0.33 ± 0.03	0.32 ± 0.02	**0.29 ± 0.02**	0.33 ± 0.01
Gene	-	-	-	-	-	0.36 ± 0.01	**0.33 ± 0.01**
Insurance	0.69 ± 0.04	0.72 ± 0.05	0.41 ± 0.09	**0.35 ± 0.05**	0.39 ± 0.08	0.39 ± 0.07	0.37 ± 0.07
Insurance3	0.45 ± 0.01	0.91 ± 0.02	0.47 ± 0.01	0.49 ± 0.02	0.41 ± 0.02	**0.45 ± 0.00**	0.45 ± 0.01
Pigs	0.95 ± 0.00	0.81 ± 0.00	-	-	-	0.56 ± 0.00	**0.51 ± 0.00**

Table 4
Time-efficiency (in seconds) comparison on the standard causal network datasets.

Datasets	MMHC	NOTEARS	DAG-GNN	PCD-by-PCD	CMB	LC-KMB$_d$	LC-KMB$_t$
Alarm	2.05	394.52	358.74	0.54	0.56	0.84	0.84
Alarm3	156.73	4936.27	5881.34	1.02	0.98	11.27	11.25
Child	1.26	74.35	92.81	0.42	0.69	0.76	0.74
Child3	107.24	1431.29	1360.78	0.31	0.30	1.98	2.01
Gene	-	-	-	-	-	32.13	36.74
Insurance	0.68	247.25	189.36	0.45	0.51	0.64	0.64
Insurance3	158.27	5431.25	7588.34	1.25	1.27	8.43	7.89
Pigs	102.75	26193.87	-	-	-	15.25	12.12

Performance Comparison: Tables 2, 3, 4 summarize the SHD, FDR, and $Time$ on the eight causal networks with 1,000 training instances, respectively. The Tables present the results in the format of $A \pm B$, where A indicates the average results and B represents the standard deviation. The best outcomes in each configuration are highlighted in bold. The symbol "-" signifies that the algorithm was unable to generate the output for the corresponding networks within an eight-hour timeframe. We can conclude from the experimental results that LC-KMB$_d$ and LC-KMB$_t$ are significantly better than the state-of-the-art local and global causal learning algorithms. Specifically, LC-KMB$_d$ and LC-KMB$_t$ outperform all comparing algorithms in terms of SHD and FDR on all datasets except Insurance. On Insurance, LC-KMB$_d$ and LC-KMB$_t$ are slightly worse than PCD-by-PCD but also achieve competitive performance. Since LC-KMB can simultaneously consider both linear and nonlinear causality, LC-KMB can retrieve more true positives in the discovered MB, which leads to a more accurate local causal structure. Compared with local causal learning methods PCD-by-PCD and CMB, LC-KMB$_d$ and LC-KMB$_t$ need to consume more time on small-scale data to identify the local causal structure, which is spent in nonlinear relation identifications. But this disadvantage reverses in large-scale data, where all other comparing algorithms can not complete the task within eight hours except LC-KMB$_d$ and LC-KMB$_t$. Additionally, LC-KMB$_d$ and LC-KMB$_t$ possess significant superiority compared with the three global learning methods, in terms of both accuracy and time efficiency, which demonstrates the practicability of the proposed methods.

4.4. Application in real-world EEG data

In this subsection, we demonstrate the effectiveness of LC-KMB in real-world emotion recognition tasks.

Dataset Description: To evaluate the effectiveness of LC-KMB in real-world applications, we applied it to identify direct causes and effects in the SEED dataset [32]. This dataset consists of electroencephalography (EEG) signals recorded by 62-channel symmetrical electrodes, with the positions of these electrodes depicted in Fig. 4(a). The dataset comprises 310 features, corresponding to five frequency bands (Delta, Theta, Alpha, Beta, and Gamma) recorded from each symmetrical electrode. The target attribute for this evaluation is the class attribute denoted as *emotions*, which includes three values: positive, neutral, and negative. The SEED dataset contains data from 15 subjects, representing the emotional data of 15 individuals, with each subject contributing 3,394 samples, consisting of 1,170 positive, 1,104 neutral, and 1,120 negative samples.

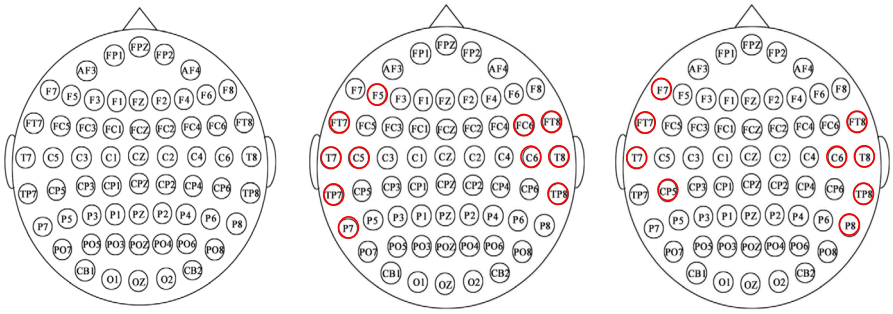

Fig. 4. (a) is the layout of 62 channel symmetrical electrodes on the EEG. (b) and (c) provide the profiles of the top 20 variables with the highest frequency of occurrences from the Beta and Gamma frequency bands. The result is consistent with previous findings in [37].

Performance Demonstration: In this experiment, we employ LC-KMB to identify the variables directly associated with the emotion recognition task among the 310 features corresponding to different channels and frequency bands. Specifically, we aim to identify the direct causes and direct effects of the target variable. To validate the effectiveness of LC-KMB, we compare the results obtained with the ESI NeuroScan System as reported in a previous study [37]. We randomly select 1000 samples from the three classes and apply LC-KMB to search for the local causal structure of the target variable. By repeating this process for each subject, we obtain 15 subsets of causal variables, which include the direct causes and effects of the target variable, *emotions*. From these subsets, we select the top 20 variables with the highest frequency of occurrences as the direct causes and effects of *emotions*. The positions of these selected variables are illustrated in Fig. 4. By analyzing Fig. 4, we can deduce that the top 20 important variables belong to two frequency bands, namely Gamma and Beta. Furthermore, these direct causes and effects are distributed in the lateral temporal area, which aligns with previous findings reported in [37]. The EEG experiment provides evidence supporting the effectiveness of LC-KMB in identifying direct causes and effects.

5. Conclusion

In this paper, we have addressed a crucial gap in existing local causal learning techniques by introducing a novel approach that overcomes limitations in identifying nonlinear and multivariate causal relationships. Our proposed method, LC-KMB, combines the power of reproducing kernel Hilbert space and the concept of conditional covariance operator to redefine local causal learning. By emphasizing the discovery of the Markov boundary (MB) through the minimization of the conditional covariance operator, we have achieved a breakthrough in capturing both linear and nonlinear causality within local causal structures.

The key strengths of our approach lie in its ability to bridge the gap between linear and nonlinear causal relationships and in its successful application to local causal structure learning. LC-KMB is not only the first algorithm to offer nonlinear local causal learning, but it also outperforms existing methods by a significant margin, as demonstrated through extensive experimentation on synthetic and real-world datasets. This performance improvement is attributed to the synergy between our kernel-based MB discovery strategy and the nonlinear Bayesian network score function, which collectively enable the identification of more accurate and realistic local causal structures.

Our contribution extends beyond the algorithm itself. By introducing the concept of kernel-based MB discovery, we provide a fresh perspective on tackling the intricate problem of local causal learning, and our successful results add a valuable dimension to the existing literature. Researchers and practitioners in the field of causal inference can benefit from our method's capacity to capture complex causal relationships in local structures. Furthermore, our work opens doors to the integration of kernel techniques into the broader causal inference landscape, inspiring the development of novel methods that merge kernel approaches with other learning paradigms.

However, we acknowledge some limitations in our work that merit discussion. While LC-KMB is a pioneering step in nonlinear local causal learning, it is not exempt from challenges. (1) The computational complexity associated with the kernel-based MB discovery strategy can still be demanding, especially for larger datasets. Additionally, although our method significantly reduces the search space by focusing on the MB, there might be situations where further optimization can be explored to enhance efficiency. (2) LC-KMB has been developed and rigorously tested within the context of well-structured data that adhere to typical assumptions of causality. However, it is not currently equipped to handle data with missing values, semi-supervised settings, imbalanced distributions, or scenarios involving specific domain characteristics or small sample sizes.

As for future work, several promising avenues are worth exploring. First, the scalability of LC-KMB could be addressed through parallel computing or approximations, enabling the application of the method to even larger datasets. Second, while we have demonstrated the effectiveness of LC-KMB for local causal structure learning, its potential for applications beyond structure discovery, such as causal effect estimation and intervention planning, should be investigated. Third, extending LC-KMB's applicability to more diverse data scenarios is an exciting avenue for future research. This involves exploring ways to adapt the method to handle missing data, accommodate unbalanced distributions, and leverage small-sample settings. Additionally, investigating techniques to incorpo-

rate domain-specific knowledge or expert constraints could enhance the algorithm's robustness and extend its usability. Given the performance demonstrated in Section 4.4, future research could apply LC-KMB to more biomedical applications [38,39] so that the performance and interpretability of learning algorithms could be promoted by causality.

CRediT authorship contribution statement

Xingyu Wu: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Yan Zhong:** Visualization, Writing – review & editing. **Zhaolong Ling:** Data curation, Resources. **Jie Yang:** Validation. **Li Li:** Writing – review & editing. **Weiguo Sheng:** Funding acquisition, Supervision. **Bingbing Jiang:** Project administration, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgements

This work was supported in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang Province (Grant No. 2023C01022), Natural Science Foundation of China (Grant No. 62006065, 62306002, 62006058, and 62366009); and in part by the Guangxi Key Laboratory of Trusted Software (Grant No. KX202311).

References

[1] Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K. Selçuk Candan, Huan Liu, Evaluation methods and measures for causal learning algorithms, IEEE Trans. Artif. Intell. 3 (6) (2022) 924–943.
[2] Ioannis Tsamardinos, Laura E. Brown, Constantin F. Aliferis, The max-min hill-climbing Bayesian network structure learning algorithm, Mach. Learn. 65 (1) (2006) 31–78.
[3] Tian Gao, Qiang Ji, Local causal discovery of direct causes and effects, Proc. Adv. Neural Inf. Process. Syst. 28 (2015).
[4] Jianxin Yin, You Zhou, Changzhang Wang, Ping He, Cheng Zheng, Zhi Geng, Partial orientation and local structural learning of causal networks for prediction, in: Causation and Prediction Challenge, PMLR, 2008, pp. 93–105.
[5] Xingyu Wu, Bingbing Jiang, Tianhao Wu, Huanhuan Chen, Practical Markov boundary learning without strong assumptions, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 10388–10398.
[6] Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, Xindong Wu, Causality-based feature selection: methods and evaluations, ACM Comput. Surv. 53 (5) (2020) 1–36.
[7] Xingyu Wu, Bingbing Jiang, Yan Zhong, Huanhuan Chen, Tolerant Markov boundary discovery for feature selection, in: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, 2020, pp. 2261–2264.
[8] Xingyu Wu, Zhenchao Tao, Bingbing Jiang, Tianhao Wu, Xin Wang, Huanhuan Chen, Domain knowledge-enhanced variable selection for biomedical data analysis, Inf. Sci. 606 (2022) 469–488.
[9] Charles R. Baker, Joint measures and cross-covariance operators, Trans. Am. Math. Soc. 186 (1973) 273–289.
[10] Kenji Fukumizu, Francis R. Bach, Michael I. Jordan, Kernel dimension reduction in regression, Ann. Stat. 37 (4) (2009) 1871–1905.
[11] Peter Spirtes, Kun Zhang, Causal Discovery and Inference: Concepts and Recent Methodological Advances, Applied Informatics, vol. 3, SpringerOpen, 2016, pp. 1–28.
[12] Raghavendra Addanki, Shiva Kasiviswanathan, Andrew McGregor, Cameron Musco, Efficient intervention design for causal discovery with latents, in: Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 63–73.
[13] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D. Ramsey, Ruben Sanchez-Romero, Clark Glymour, Bernhard Schölkopf, Causal discovery from heterogeneous/nonstationary data, J. Mach. Learn. Res. 21 (89) (2020) 1–53.
[14] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, Huan Liu, A survey of learning causality with data: problems and methods, ACM Comput. Surv. 53 (4) (2020) 1–37.
[15] David Maxwell Chickering, Optimal structure identification with greedy search, J. Mach. Learn. Res. 3 (11) (2002) 507–554.
[16] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, Antti Kerminen, Michael Jordan, A linear non-Gaussian acyclic model for causal discovery, J. Mach. Learn. Res. 7 (10) (2006).
[17] Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, Eric P. Xing, Dags with no tears: continuous optimization for structure learning, Proc. Adv. Neural Inf. Process. Syst. 31 (2018).
[18] Shuai Yang, Hao Wang, Kui Yu, Fuyuan Cao, Xindong Wu, Towards efficient local causal structure learning, in: IEEE Transactions on Big Data, 2021.
[19] Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, Xenofon D. Koutsoukos, Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation, J. Mach. Learn. Res. 11 (1) (2010) 171–234.
[20] Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, Xenofon D. Koutsoukos, Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions, J. Mach. Learn. Res. 11 (1) (2010) 235–284.
[21] Isabelle Guyon, Constantin Aliferis, et al., Causal feature selection, in: Computational Methods of Feature Selection, Chapman and Hall/CRC, 2007, pp. 79–102.
[22] Xingyu Wu, Bingbing Jiang, Yan Zhong, Huanhuan Chen, Multi-target Markov boundary discovery: theory, algorithm, and application, IEEE Trans. Pattern Anal. Mach. Intell. 45 (4) (2022) 4964–4980.
[23] Xingyu Wu, Bingbing Jiang, Xiangyu Wang, Taiyu Ban, Huanhuan Chen, Feature selection in the data stream based on incremental Markov boundary learning, IEEE Trans. Neural Netw. Learn. Syst. 34 (10) (2023) 6740–6754.

[24] Niinimaki Teppo, Pekka Parviainen, Local structure discovery in Bayesian networks, in: Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence, 2012, pp. 634–643.
[25] Tian Gao, Qiang Ji, Efficient score-based Markov blanket discovery, Int. J. Approx. Reason. 80 (2017) 277–293.
[26] Zhifa Liu, Brandon Malone, Changhe Yuan, Empirical evaluation of scoring functions for Bayesian network model selection 13 (15) (2012) 1–16.
[27] Xingyu Wu, Bingbing Jiang, Kui Yu, Huanhuan Chen, et al., Accurate Markov boundary discovery for causal feature selection, IEEE Trans. Cybern. 50 (12) (2020) 4983–4996.
[28] Xingyu Wu, Bingbing Jiang, Kui Yu, Huanhuan Chen, Separation and recovery Markov boundary discovery and its application in eeg-based emotion recognition, Inf. Sci. 571 (2021) 262–278.
[29] Richard S. Hamilton, Three-manifolds with positive Ricci curvature, J. Differ. Geom. 17 (2) (1982) 255–306.
[30] Gene H. Golub, Per Christian Hansen, Dianne P. O'Leary, Tikhonov regularization and total least squares, SIAM J. Matrix Anal. Appl. 21 (1) (1999) 185–194.
[31] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, Clark Glymour, Generalized score functions for causal discovery, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1551–1560.
[32] Ruo-Nan Duan, Jia-Yi Zhu, Bao-Liang Lu, Differential entropy feature for EEG-based emotion classification, in: Proceedings of the 6th IEEE International Conference on Neural Engineering, 2013, pp. 81–84.
[33] Ioannis Tsamardinos, Constantin F. Aliferis, Alexander R. Statnikov, Er Statnikov, Algorithms for large scale Markov blanket discovery, in: Proceedings of the Florida Artificial Intelligence Research Society Conference, 2003, pp. 376–380.
[34] Constantin F. Aliferis, Ioannis Tsamardinos, Alexander Statnikov, HITON: a novel Markov blanket algorithm for optimal variable selection, in: Proceedings of the American Medical Informatics Association Annual Symposium, 2003, pp. 21–25.
[35] John H. McDonald, Handbook of Biological Statistics, vol. 2, Sparky House Publishing, Baltimore, MD, 2009.
[36] Alexander Statnikov, Constantin F. Aliferis Tied, An artificially simulated dataset with multiple Markov boundaries, in: Journal of Machine Learning Research Workshop and Conference Proceedings, Volume 6: Causality: Objectives and Assessment, 2010, pp. 249–256.
[37] Wei-Long Zheng, Bao-Liang Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, IEEE Trans. Auton. Ment. Dev. 7 (3) (2015) 162–175.
[38] Jyotismita Chaki, Marcin Woźniak, A deep learning based four-fold approach to classify brain mri: Btscnet, Biomed. Signal Process. Control 85 (2023) 104902–104922.
[39] Marcin Woźniak, Jakub Siłka, Michał Wieczorek, Deep neural network correlation learning mechanism for ct brain tumor detection, Neural Comput. Appl. (2021) 1–16.