Statistica Sinica **35** (2025), 479-504
doi:https://doi.org/10.5705/ss.202022.0272

EFFICIENT LEARNING OF NONPARAMETRIC DIRECTED ACYCLIC GRAPH WITH STATISTICAL GUARANTEE

Yibo Deng[1], Xin He[1] and Shaogao Lv*[2]

[1]*Shanghai University of Finance and Economics*
and [2]*Nanjing University of Aeronautics and Astronautics*

Abstract: Directed acyclic graph (DAG) models are widely used to represent casual relations among collected nodes. This paper proposes an efficient and consistent method to learn DAG with a general causal dependence structure, which is in sharp contrast to most existing methods assuming linear dependence of causal relations. To facilitate DAG learning, the proposed method leverages the concept of topological layer, and connects nonparametric DAG learning with kernel ridge regression in a smooth reproducing kernel Hilbert space (RKHS) and learning gradients by showing that the topological layers of a nonparametric DAG can be exactly reconstructed via kernel-based estimation, and the parent-child relations can be obtained directly by computing the estimated gradient function. The developed algorithm is computationally efficient in the sense that it attempts to solve a convex optimization problem with an analytic solution, and the gradient functions can be directly computed by using the derivative reproducing property in the smooth RKHS. The asymptotic properties of the proposed method are established in terms of exact DAG recovery without requiring any explicit model specification. Its superior performance is also supported by a variety of simulated and a real-life example.

Key words and phrases: Causality, exact DAG recovery, learning gradients, nonparametric DAG, RKHS.

1. Introduction

Directed acyclic graph (DAG) models are widely used to represent directional or parent-child relations among interacting units, which have a wide range of applications in many disciplines (Spirtes, Glymour and Scheines, 2000; Peters, Janzing and Schölkopf, 2017). Thus, learning DAG from the observed data has attracted tremendous attention in the past decades (Shimizu et al., 2011; Peters and Bühlmann, 2014; Yuan et al., 2019; Zhao, He and Wang, 2022) and is still challenging especially when the casual relations display a general dependence structure beyond linearity (Bühlmann, Peters and Ernest, 2013; Peters et al., 2014; Park, 2020; Gao, Ding and Aragam, 2020).

*Corresponding author. E-mail: lvsg716@nau.edu.cn

In literature, most existing DAG learning methods assume that the parent-child relations have a linear dependence structure and thus assume the linear structural equation models (SEMs). These methods can be roughly categorized into three classes. The first class attempts to learn linear Gaussian DAG by assuming that all the noise terms are Gaussian distributed. Specifically, Peters and Bühlmann (2014) shows that a linear Gaussian DAG is identifiable if all the noise terms have equal variances, and then motivates a variety of learning methods (Yuan et al., 2019; Chen, Drton and Wang, 2019; Li, Shen and Pan, 2020). The second class focuses on learning linear non-Gaussian DAG. One of the most important works is that Shimizu, Hyvärinen and Kerminen (2006) proves that a linear non-Gaussian DAG is identifiable if all the noise terms follow continuous non-Gaussian distribution, and an iterative search algorithm is developed. This fundamental work also motivates a variety of follow-up studies (Shimizu et al., 2011; Hyvärinen and Smith, 2013; Wang and Drton, 2020; Zhao, He and Wang, 2022). Recently, Park and Raskutti (2018) and Zhou et al. (2022) focus on a general class of non-Gaussian DAG models that the conditional variance of each node given its parents is a quadratic function of its conditional mean, which admits many non-Gaussian distributions including some discrete ones. The other class of methods further relaxes the distribution assumption by requiring some explicit order among noise variances (Ghoshal and Honorio, 2018; Park, 2020). Note that almost all the methods in these categories are designed to recover causal relations with linear dependence structure. Yet, as pointed out by Yuan et al. (2019), many causal relations in real-life analysis may have nonlinear behavior that cannot be captured by any linear model.

Nonparametric DAG relaxes the linear dependence assumption by allowing more general causal relations, and thus has attracted tremendous interest in recent years (Bühlmann, Peters and Ernest, 2013; Peters et al., 2014; Mooij et al., 2016; Rothenhäusler, Ernest and Bühlmann, 2018; Park, 2020; Zhang et al., 2020; Gao, Ding and Aragam, 2020; Li, Shen and Pan, 2023). A majority class of learning nonparametric DAG methods replace the linear SEMs with the additive noise models (ANMs), where each node is generated by a nonparametric function of its parents adding an independent noise term. Moreover, additive modelling (Stone, 1985) is often imposed to model the nonparametric function. Specifically, Bühlmann, Peters and Ernest (2013) proposes a casual additive model and aims to learn the DAG via maximum likelihood estimation and variable selection technique for additive modelling. Peters et al. (2014) proposes the RESIT algorithm to learn a potential causal ordering via sequential nonparametric fitting and independence testing. Rothenhäusler, Ernest and Bühlmann (2018) further considers the case that the nonparametric function is a partially linear model under the additive modelling assumption. Some other classes of methods focus on the bivariate models or the post-nonlinear models (Zhang and Hyvärinen, 2009; Zhang et al., 2016) and the score-based search procedures within a more

general function space of the nonparametric function (Zhang et al., 2020). It should be pointed out that all the aforementioned methods attempt to recover an indeterministic causal ordering, and many of them lack theoretical guarantee in terms of exact DAG recovery or suffer computational burden even when dealing with a medium-sized DAG.

Most recently, Gao, Ding and Aragam (2020) introduces the concept of layers into nonparametric DAG learning to eliminate the unnecessary inefficiency caused by casual ordering. Specifically, it estimates the nonparametric function with some standard nonparametric estimators, including the kernel smoother, nearest neighbors, and additive modeling with splines, to recover the layer structure, and then adopts the variable selection technique for additive modeling to recover the parent-child relations after all the layers being estimated. Note that the recovery procedure may suffer computational burden when the number of nodes is relatively large, even using the additive modeling with splines, not to mention the kernel smoother or nearest neighbor estimator. Moreover, their proposed method is mainly designed for the special case with equal variances, and their theoretical analysis only focuses on establishing the layer recovery consistency by assuming that the employed nonparametric estimator is consistent. Yet, the statistical guarantee in terms of exact DAG recovery remains largely unknown especially when the employed additive modeling assumption is violated.

In this paper, we propose an efficient method to learn nonparametric DAG with theoretical guarantee. A useful concept of topological layer is adopted to facilitate DAG learning, which assures that any DAG can be converted into a unique topological structure, where the parents of a node must belong to its upper layers, and thus acyclicity is naturally guaranteed. The proposed method is motivated by the key fact that topological layers of a nonparametric DAG with heterogeneous noise variances are identifiable, and the general parent-child relation can be fully detected by gradient functions. The proposed method adopts kernel-based estimation in the RKHS for reconstructing layers and the parent-child relations can be simultaneously recovered as a by-product via learning gradients. The proposed method is computationally efficient and its asymptotic properties are provided in terms of exact DAG recovery, which are established without requiring any specific model assumption. Its superior performance is also supported by a variety of simulated and real-life examples.

The main contribution of this paper is the development of an efficient learning method to learn nonparametric DAG from observed data, and the investigation of its statistical guarantees in terms of exact DAG recovery. Specifically, we show that the topological layers of a nonparametric DAG can be sequentially reconstructed under the conditional noise variance assumption in a top-down fashion, and the gradient function can be employed as a useful tool to recover the general parent-child relations. More importantly, we connect nonparametric DAG learning with kernel ridge regression and learning gradients by showing

that the layers can be exactly reconstructed via kernel-based estimation, and the parent-child relations can be simultaneously obtained by computing the estimated gradient function without any extra estimation. Computationally, an efficient learning algorithm is developed, where the corresponding convex optimization task has an analytic solution, and the derivative reproducing property in RKHS ensures that the gradient function can be directly computed. Theoretically, with the help of functional operators in learning theory, the statistical guarantees of the proposed method are established ensuring the underlying DAG with general parent-child relations can be exactly recovered, which is particularly attractive in line of research in nonparametric DAG learning.

The rest of this paper is organized as follows. Section 2 introduces some background of nonparametric DAG, the concept of topological layers, and the motivations of the proposed method. Section 3 develops an efficient algorithm for learning nonparametric DAG, and Section 4 establishes the theoretical results of the proposed method in terms of exact DAG recovery under mild conditions. Numerical experiments on several simulated examples and a real-life analysis are provided in Section 5. Section 6 contains a brief discussion, and all the technical proofs are provided in an online supplementary file.

2. Learning Nonparametric Directed Acyclic Graph

We consider a directed acyclic graph (DAG) model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ encoding the joint distribution $P(\mathbf{x})$ of variables $\mathbf{x} = (x_1, \ldots, x_p)^T \in \mathcal{X} \subset \mathcal{R}^p$. Precisely, $\mathcal{V} = \{1, \ldots, p\}$ represents a set of nodes associated with $\mathbf{x}$, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ denotes a set of directed edges without directed cycles representing the parent-child relations. For notation ease, we denote node k's parents as $\mathrm{pa}_k \subset \{1, \ldots, p\}$ and its non-descendants (exclude itself) by nd_k. For any $j \in \mathrm{pa}_k$, an arrow from x_j towards x_k in $\mathcal{G}$ is indicated; if x_k has no parents in $\mathcal{G}$, such as x_k is a root or isolated node, we have $\mathrm{pa}_k = \emptyset$. Moreover, we denote the set of all the directed edges pointing to node k as $\mathcal{E}_k = \{j \to k, \text{ for any } j \in \mathrm{pa}_k\}$. We also assume that the Markov property (Spirtes, Glymour and Scheines, 2000; Yuan et al., 2019) and causal minimality (Bühlmann, Peters and Ernest, 2013) hold. To be more precise, the Markov property requires that $P(\mathbf{x})$ can be factorized based on $\mathcal{G}$ into the product of the conditional distributions of each variable given their parents that $P(\mathbf{x}) = \prod_{k=1}^{p} P(x_k | \mathbf{x}_{\mathrm{pa}_k})$, where $\mathbf{x}_{\mathrm{pa}_k}$ denotes all the variables x_j, $j \in \mathrm{pa}_k$.

To represent the causal structure, we apply the continuous additive noise model (ANM) which is also known as the functional model (Peters et al., 2014). Note that ANMs are a special case of DAG models where the joint distribution is defined by the following structural equations with additive noise. Precisely, each x_j is centered with mean zero, the graph structure can be represented by

the following ANM that

$$x_j = f_j^*(\mathbf{x}_{\mathrm{pa}_j}) + n_j, \quad \text{for any} \quad j \in \mathcal{V}, \tag{2.1}$$

where $f_j^*(\mathbf{x}_{\mathrm{pa}_j}) = E(x_j|\mathbf{x}_{\mathrm{pa}_j}), j \in \mathcal{V}$ are allowed to have any form of Borel measurable functions and are assumed to be differentiable, and the noise terms $\{n_j\}_{j\in\mathcal{V}}$ have strictly positive densities and are independent but may allow following different distributions with mean zero and heterogeneous variances that $E(n_j) = 0$ and $\mathrm{Var}(n_j) = \sigma_j^2$. This is much more general than most existing works which either assume f_j^* following linear (Ghoshal and Honorio, 2018; Yuan et al., 2019) or additive model assumption (Bühlmann, Peters and Ernest, 2013; Gao, Ding and Aragam, 2020). It is worthy pointing out that the requirement that each x_j is centered with mean zero and $E(n_j) = 0$ imply that the true target function f_j^* in (2.1) also has zero mean.

We now introduce the RKHS $\mathcal{H}_K$ associated with a specified kernel K taking values on a subset of $\mathcal{R}^p$ and endowed with the norm $\|\cdot\|_K$. It is well-known that RKHS induced by some universal kernel, such as the Gaussian kernel, is differentiable and fairly large in the sense that any continuous function can be well approximated by some intermediate function in the induced RKHS under the infinity norm (Steinwart and Christmann, 2008). To be more precise, we have $K_{\mathbf{x}} := K(\mathbf{x}, \cdot) \in \mathcal{H}_K$ for any $\mathbf{x} \in \mathcal{X}$, and $\langle f, K_{\mathbf{x}}\rangle_K = f(\mathbf{x})$ for any $f \in \mathcal{H}_K$. By the Mercer's theorem (Steinwart and Christmann, 2008), under some regularity conditions, the eigen-expansion of the kernel function is

$$K(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{\infty} \mu_k \phi_k(\mathbf{x})\phi_k(\mathbf{x}'), \quad \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X},$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are non-negative eigenvalues, and $\{\phi_k\}_{k=1}^{\infty}$ are the associated eigenfunctions, taken to be orthonormal in $\mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}}) = \{f : \int_{\mathcal{X}} f(\mathbf{x})^2 d\rho_{\mathbf{x}} < \infty\}$ with $\rho_{\mathbf{x}}$ denoting the marginal distribution of $\mathbf{x}$. Moreover, the RKHS-norm of any $f \in \mathcal{H}_K$ then can be written as $\|f\|_K^2 = \sum_{k\geq 1} a_k^2/\mu_k$ where $a_k = \langle f, \phi_k\rangle_{\mathcal{L}^2(\mathcal{X},\rho_{\mathbf{x}})}^2 = \int_{\mathcal{X}} f(\mathbf{x})\phi_k(\mathbf{x}) d\rho_{\mathbf{x}}$ denote Fourier coefficients, and thus for any $f \in \mathcal{H}_K$, we have $f(\mathbf{x}) = \sum_{k=1}^{\infty} a_k \phi_k(\mathbf{x})$. Note that these results require that $\mathcal{H}_K \subset \mathcal{L}^2(\mathcal{X}, \rho_{\mathbf{x}})$, which is automatically satisfied if $\sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x}, \mathbf{x})$ is bounded. Then, the RKHS induced by the kernel K can be written as

$$\mathcal{H}_{K,p} := \left\{ f = \sum_{k=1}^{\infty} a_k \phi_k \;\middle|\; \sum_{k\geq 1} \frac{a_k^2}{\mu_k} \leq \infty \right\}.$$

It is important to notice that in the rest of this paper, we need to search functions sequentially over different RKHS induced by the kernel function with different inputs' dimensions. With a slight abuse of notation, we write all the RKHSs as $\mathcal{H}_K$ when the inputs' dimension of the corresponding kernel function is clear for

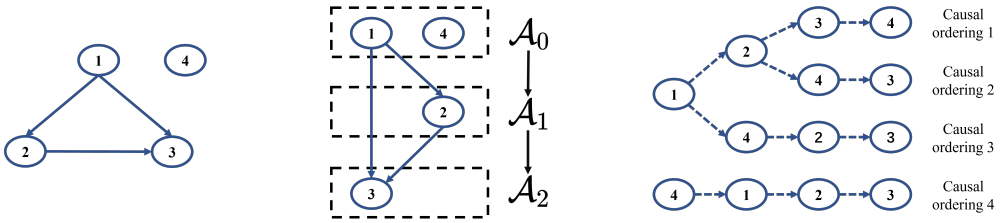


Figure 1. An illustration of a DAG with 3 layers.

notation simplicity. Moreover, we assume that $Ef(\mathbf{x}) = \int_{\mathcal{X}} f(\mathbf{x}) d\rho_{\mathbf{x}} = 0$ for all $f \in \mathcal{H}_K$ to facilitate DAG learning. It is worthy pointing out that this zero mean assumption is also required in many kernel-based learning problems and can be verified if the kernel function is centralized (He et al., 2022). In literature, various centralized kernels have been proposed, including the centralized Gaussian kernel and interested readers are referred to Lindsay et al. (2008), Durrande et al. (2013) and He et al. (2022) for detailed discussions.

2.1. Learning with topological layers

Without loss of generality, we assume that a DAG $\mathcal{G}$ has T layers, for some positive constant T, and each node only belongs to one layer, due to its longest path to roots (nodes with no parents). Note that the concept of a topological layer is explicitly defined in Zhao, He and Wang (2022) and Zhou et al. (2022) for learning linear DAGs, which is general and which is general and reconstructs any DAG in such a way that causal ordering among each layer is uniquely determined, and a similar idea is also adopted to learn nonparametric DAG in Gao, Ding and Aragam (2020) and Li, Shen and Pan (2023). Note that the idea of topological layers significantly differs from the commonly used causal ordering in literature (Yuan et al., 2019). More importantly, we show that the procedure of learning nonparametric DAG can be much more stable and computationally efficient, and can establish theoretical guarantees in terms of exact DAG recovery.

Specifically, let $\mathcal{A}_t$ denote all the nodes in the t-th layer and $\mathcal{S}_t = \cup_{d=0}^{t-1} \mathcal{A}_d$ denote the nodes in all the upper layers. Clearly, we have $\mathcal{S}_0 = \emptyset$ and $\mathcal{S}_T = \mathcal{V}$. Figure 1 illustrates a toy DAG with its unique topological layer structure.

From Figure 1, we see that nodes 1 and 4 are regarded as root and isolated node, respectively, and thus belong to the first layer $\mathcal{A}_0$; although node 1 is one of its parents, node 3 still belongs to the last layer $\mathcal{A}_2$ due to its longest path to root ($1 \to 2 \to 3$). It is worth noting that node 4 is named as an isolated node due to the fact that it does not direct to any other nodes. In fact, node 4 can also be regarded as a root which has no children. In sharp contrast, the toy example has multiple potential causal orderings as illustrated in the left panel of Figure 1, which may lead to unnecessary estimation instability and computational inefficiency in recovering the DAG

structures. Clearly, any nonparametric DAG can be uniquely converted to the corresponding topological structure, thus the original task of DAG learning can be decomposed into reconstructing the layers and recovering the parent-child relations among layers.

2.2. Reconstruction via topological layer

In this section, we show that the topological layer of a nonparametric DAG can be reconstructed under mild conditions, and the causal minimality condition connects the recovery of parent-child relations with learning gradients. Let de_j denote all the descendant nodes of node j, and then the topological layers of a nonparametric DAG model can be identified under the following technical condition.

Assumption 1. *For any $j, j' \in \mathcal{A}_t$, $t = 0, \ldots, T-1$ and $k \in de_j$, there exists some positive quantity $M_{\max}$ such that $\min_{j,k} \sigma_k^2 + E[\mathrm{Var}\{E(x_k|\mathbf{x}_{pa_k})|\mathbf{x}_{\mathcal{S}_t}\}] - \sigma_j^2 > M_{\max}$ and $\sigma_j^2 = \sigma_{j'}^2$.*

Assumption 1 is a general condition and is widely used in literature of learning ANMs (Park, 2020). Particularly, the first part of Assumption 1 allows that nodes belonging to different layers have heterogeneous variance, which relaxes the commonly used equal variance assumption (Gao, Ding and Aragam, 2020), and is analogous to the conditions required in Theorem 2 of Park (2020) in terms of causal ordering. The second part of Assumption 1 requires that nodes belonging to the same layer have equal variance, which is natural in the sense that they may come from a similar domain, and thus share a similar characteristic. Note that the equal variance condition can be further relaxed by allowing nodes in the same layer to have heterogeneous variances, but their differences are upper bounded by some constant less than $M_{\max}$.

Theorem 1. *Consider an ANM (2.1) associated with DAG $\mathcal{G}$. Suppose that $\mathcal{A}_0, \ldots, \mathcal{A}_{t-1}$ have been identified and $\mathcal{S}_t = \cup_{d=0}^{t-1} \mathcal{A}_d$. Then, for any $t = 0, \ldots, T-1$, there holds*

$$E\big\{\mathrm{Var}(x_j|\mathbf{x}_{\mathcal{S}_t})\big\} = \begin{cases} \sigma_j^2, & \text{for any } j \in \mathcal{A}_t; \\ \sigma_j^2 + E\big[\mathrm{Var}\big\{E(x_j|\mathbf{x}_{pa_j})|\mathbf{x}_{\mathcal{S}_t}\big\}\big], & \text{for any } j \in \mathcal{V}\backslash\{\mathcal{S}_t \cup \mathcal{A}_t\}. \end{cases} \tag{2.2}$$

Additionally, suppose that Assumption 1 is satisfied, then the topological layers can be exactly reconstructed.

Theorem 1 ensures that the topological layers can be reconstructed in a hierarchical fashion by evaluating the conditional variance for each remaining node. The first part of Theorem 1 states that if node j belongs to the current layer $\mathcal{A}_t$, the expected conditional variance is exactly the same as the corresponding

noise variance; otherwise, the expected conditional variance should be strictly larger than the noise variance. Moreover, by assuming Assumption 1, the second part of Theorem 1 ensures that the expected conditional variances of nodes belonging to the current layer are exactly the same, and there exists some gap between the expected conditional variances of nodes belonging to the current layer and to all the lower layers. And thus, the topological layer can be reconstructed correspondingly. Particularly, Theorem 1 shows that for any $j \in \mathcal{A}_0$ with $\mathrm{pa}_j = \emptyset$, the layer $\mathcal{A}_0$ can be exactly reconstructed by the fact that $\mathrm{Var}(x_j) = \sigma^2_{0,\min}$ for any $j \in \mathcal{A}_0$ and otherwise $\mathrm{Var}(x_j) > \sigma^2_{0,\min}$ with $\sigma^2_{0,\min} = \min_{k\in\mathcal{V}} \mathrm{Var}(x_k)$. For the general cases that $t = 1, \ldots, T-1$, Theorem 1 ensures that for any $j \in \mathcal{A}_t$, there holds $E\{\mathrm{Var}(x_j|\mathbf{x}_{\mathcal{S}_t})\} = \sigma^2_{t,\min}$ with $\sigma^2_{t,\min} = \min_{k\in\mathcal{V}\setminus\{\mathcal{S}_t\}} E\{\mathrm{Var}(x_k|\mathbf{x}_{\mathcal{S}_t})\}$, and $E\{\mathrm{Var}(x_\ell|\mathbf{x}_{\mathcal{S}_t})\} > \sigma^2_{t,\min} + M_{\max}$ for any node belonging to lower layers. Notably, Theorem 1 ensures that the layers can be reconstructed in a top-down fashion, whereas Theorem 2 of Park (2020) shows that causal ordering can be forward or backward recovered under different types of noise-variance assumptions. In fact, Theorem 1 as well as our motivated method can be further extended to reconstruct the topological layers in the bottom-up fashion and more discussions on this possible extension are provided in Section 6.

More interestingly, among the above reconstruction procedures, suppose that $\mathcal{A}_0, \ldots, \mathcal{A}_t$ have been identified. By the definition of $\mathcal{A}_t$, for any node $j \in \mathcal{A}_t$, we have $\mathrm{pa}_j \subset \mathcal{S}_t = \cup_{d=0}^{t-1}\mathcal{A}_d$ and $\mathrm{de}_j \cap \mathcal{S}_t = \emptyset$, and thus there holds $f^*_j(\mathbf{x}_{\mathrm{pa}_j}) = E(x_j|\mathbf{x}_{\mathrm{pa}_j}) = E(x_j|\mathbf{x}_{\mathcal{S}_t}) = f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t})$. Furthermore, we notice that as pointed out in Section 3 of Peters et al. (2014), causal minimality reduces to the condition that each function f^*_j is not constant in any of its arguments under (2.1). This requires that all the parents should make a contribution to their child, and implies pa_j is the set of nodes with non-zero gradients. Precisely, by assuming causal minimality, for any $j \in \mathcal{A}_t$, we have

$$\|g^*_{jk}\|_2^2 = \int \left\{ \frac{\partial f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t})}{\partial x_k} \right\}^2 d\rho_{\mathbf{x}_{\mathcal{S}_t}} > 0, \quad \text{for any} \quad k \in \mathrm{pa}_j, \tag{2.3}$$

and for any $k \in \mathcal{S}_t \setminus \{\mathrm{pa}_j\}$, there holds $\|g^*_{jk}\|_2^2 = 0$, due to the fact that $f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t}) = E(x_j|\mathbf{x}_{\mathcal{S}_t}) = E(x_j|\mathbf{x}_{\mathrm{pa}_j})$ since $j \in \mathcal{A}_t$ and in (2.3), each gradient function is evaluated given all the other nodes belonging to pa_j. Thus, for any node $j \in \mathcal{A}_t$, pa_j can be written as $\mathrm{pa}_j = \{k \in \mathcal{S}_t, \|g^*_{jk}\|_2^2 > 0\}$. It is also interesting to notice that in Section 4, the minimal signal strength is required in Assumption 4 to establish the asymptotic consistency under the finite sample setting.

Theorem 2. *Suppose that all the assumptions in Theorem* 1 *are satisfied and the causal minimality holds. Then, the DAG $\mathcal{G}$ is uniquely identifiable.*

Theorem 2 provides the identifiability result of the nonparametric DAG under the noise variance condition in Assumption 1. Its proof directly follows from Theorem 1 that all the topological layers can be exactly recovered by evaluating

the conditional variances, and from the required causal minimality assumption that once the layers are exactly identified, the underlying parent-child relations can be exactly recovered by checking the corresponding gradient functions. Therefore, we omit its proof here. Note that the established identifiability results differ from the classical identifiable results of nonparametric in literature (Peters et al., 2014) that they are motivated by different identifiable conditions. Specifically, the results in Theorem 2 do not require the noise terms to be normally distributed (Peters et al., 2014; Li, Shen and Pan, 2023) but require the noise terms have some ordered variances as stated in Assumption 1. Crucially, we notice that by the derivative reproducing property (Zhou, 2007), there holds

$$g_{jk}(\mathbf{x}) = \frac{\partial f_j(\mathbf{x})}{\partial x_k} = \langle f_j, \partial_k K_{\mathbf{x}} \rangle_K \leq \|\partial_k K_{\mathbf{x}}\|_K \|f_j\|_K, \tag{2.4}$$

for any $f_j \in \mathcal{H}_K$ and $\partial_k K_{\mathbf{x}} = \partial K(\mathbf{x}, \cdot)/\partial x_k$. This implies that the gradient function of any $f_j \in \mathcal{H}_K$ can be bounded by its K-norm up to some constant. In other words, if we want to estimate $g_{jk}(\mathbf{x})$ within the smooth RKHS, it suffices to estimate f_j itself without loss of information. Most importantly, the key factor (2.4) ensures us that the corresponding gradient function can be directly obtained if the estimator of f_j belonging to $\mathcal{H}_K$ is provided, and thus the parent-child relations can be simultaneously obtained without any extra estimation. Due to the nice properties of $\mathcal{H}_K$, we consider the estimation procedures in the smooth RKHS in the next section.

3. Nonparametric DAG Learning Algorithm

In this section, we develop an efficient learning algorithm, which connects nonparametric DAG learning and learning gradients in the smooth RKHS. Particularly, motivated by Theorem 1 and the key factor (2.3), the problem of learning a nonparametric DAG can be decomposed into a hierarchical procedure, where the topological layers can be reconstructed by computing the criteria of Theorem 1 in a top-down fashion, and simultaneously the directed edges can be directly recovered using the computed gradients in a parallel fashion.

3.1. Proposed algorithm

Given a random sample $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \in \mathcal{R}^{n\times p}$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is generated from Model (2.1) and x_{ij} denotes the i-th observation of x_j, we first attempt to reconstruct the first layer $\mathcal{A}_0$ from the observed data. Specifically, for each $j \in \mathcal{V}$, we compute the unconditional variance that $\widehat{\mathrm{Var}}(x_j) = \{n/(n-1)\}[\widehat{E}(x_j^2) - \{\widehat{E}(x_j)\}^2]$ with $\widehat{E}(x_j^2) = \sum_{i=1}^n x_{ij}^2/n$ and $\widehat{E}(x_j) = \sum_{i=1}^n x_{ij}/n$. Then, by Theorem 1, the first layer can be reconstructed as $\widehat{\mathcal{A}}_0 = \{k, |\widehat{\mathrm{Var}}(x_k) - \widehat{\sigma}_{\min}^{(0)}| < \epsilon_0\}$ with $\widehat{\sigma}_{\min}^{(0)} = \min_{j\in\mathcal{V}} \widehat{\mathrm{Var}}(x_j)$ for some small $\epsilon_0 > 0$.

Suppose that the layers $\widehat{\mathcal{A}}_0, \ldots, \widehat{\mathcal{A}}_{t-1}$ have been reconstructed and denote $\widehat{\mathcal{S}}_t = \cup_{d=0}^{t-1} \widehat{\mathcal{A}}_d$. Then, we turn to reconstruct the layer $\mathcal{A}_t$ by calculating the criteria in Theorem 1 based on the remaining nodes. Specifically, given $\widehat{\mathcal{S}}_t$, for any $j \in \mathcal{V} \backslash \{\widehat{\mathcal{S}}_t\}$, we compute the estimated criteria as

$$\widehat{E}\{\widehat{\mathrm{Var}}(x_j|\mathbf{x}_{\widehat{\mathcal{S}}_t})\} = \widehat{E}(x_j^2) - \widehat{E}\{\widehat{E}(x_j|\mathbf{x}_{\widehat{\mathcal{S}}_t})^2\}, \tag{3.1}$$

where $\widehat{E}(x_j^2) = \sum_{i=1}^n x_{ij}^2/n$ and $\widehat{E}\{\widehat{E}(x_j|\mathbf{x}_{\widehat{\mathcal{S}}_t})^2\}$ can be obtained by kernel ridge estimation in the smooth RKHS. Put differently, for each $j \in \mathcal{V} \backslash \{\widehat{\mathcal{S}}_t\}$, we regress x_j on $\mathbf{x}_{\widehat{\mathcal{S}}_t}$ by fitting a kernel ridge regression that

$$\widehat{f}_j = \underset{f_j \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left\{x_{ij} - f_j(\mathbf{x}_{i\widehat{\mathcal{S}}_t})\right\}^2 + \lambda \|f_j\|_K^2. \tag{3.2}$$

It is clear that $\widehat{f}_j(\mathbf{x}_{\widehat{\mathcal{S}}_t})$ can be treated as a valid estimation of $E(x_j|\mathbf{x}_{\widehat{\mathcal{S}}_t})$, and thus the second term in (3.1) can be computed as $\widehat{E}\{\widehat{E}(x_j|\mathbf{x}_{\widehat{\mathcal{S}}_t})^2\} = \sum_{i=1}^n \{\widehat{f}_j(\mathbf{x}_{i\widehat{\mathcal{S}}_t})\}^2/n$. Note that the employed estimation procedure (3.2) is computationally efficient and by the representer theorem (Wahba, 1998), the minimizer of (3.2) must have the following form that

$$\widehat{f}_j(\mathbf{x}_{\widehat{\mathcal{S}}_t}) = \sum_{i=1}^n \widehat{\alpha}_i^{(j)} K(\mathbf{x}_{i\widehat{\mathcal{S}}_t}, \mathbf{x}_{\widehat{\mathcal{S}}_t}) = \widehat{\boldsymbol{\alpha}}_j^T \mathbf{K}_n(\mathbf{x}_{\widehat{\mathcal{S}}_t}),$$

where $\widehat{\boldsymbol{\alpha}}_j = (\widehat{\alpha}_1^{(j)}, \ldots, \widehat{\alpha}_n^{(j)})^T$ and $\mathbf{K}_n(\mathbf{x}_{\widehat{\mathcal{S}}_t}) = \big(K(\mathbf{x}_{1\widehat{\mathcal{S}}_t}, \mathbf{x}_{\widehat{\mathcal{S}}_t}), \ldots, K(\mathbf{x}_{n\widehat{\mathcal{S}}_t}, \mathbf{x}_{\widehat{\mathcal{S}}_t})\big)^T$. Therefore, the optimization problem (3.2) has an analytic solution that $\widehat{\boldsymbol{\alpha}}_j = \big(\mathbf{K}_{\widehat{\mathcal{S}}_t}^T \mathbf{K}_{\widehat{\mathcal{S}}_t} + n\lambda \mathbf{K}_{\widehat{\mathcal{S}}_t}\big)^+ \mathbf{K}_{\widehat{\mathcal{S}}_t}^T \mathbf{x}_j$ where $\mathbf{K}_{\widehat{\mathcal{S}}_t} = \{K(\mathbf{x}_{i\widehat{\mathcal{S}}_t}, \mathbf{x}_{j\widehat{\mathcal{S}}_t})\}_{i,j=1}^n \in \mathcal{R}^{n\times n}$ denote the kernel matrix and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$. Then, by Theorem 1, the layer $\mathcal{A}_t$ can be reconstructed as $\widehat{\mathcal{A}}_t = \big\{k, |\widehat{E\mathrm{Var}}(x_k|\mathbf{x}_{\widehat{\mathcal{S}}_t}) - \widehat{\sigma}_{\min}^{(t)}| < \epsilon_t\big\}$ with $\widehat{\sigma}_{\min}^{(t)} = \min_{j\in\mathcal{V}\backslash\widehat{\mathcal{S}}_t} \widehat{E\mathrm{Var}}(x_j|\mathbf{x}_{\widehat{\mathcal{S}}_t})$ and for some small $\epsilon_t > 0$.

Once $\widehat{\mathcal{A}}_t$ is reconstructed, the parent-child relations among nodes in $\widehat{\mathcal{A}}_t$ and $\widehat{\mathcal{S}}_t$ can be simultaneously recovered by using the derivative reproducing property (2.4) as a by-product. Specifically, for each $j \in \widehat{\mathcal{A}}_t$ and $k \in \widehat{\mathcal{S}}_t$, we compute the corresponding gradient function and evaluate the existence of a directed edge by using the empirical norm that

$$\|\widehat{g}_{jk}\|_n^2 = \frac{1}{n} \sum_{i=1}^n \left\{\widehat{g}_{jk}(\mathbf{x}_{i\widehat{\mathcal{S}}_t})\right\}^2 = \frac{1}{n} \sum_{i=1}^n \left\{\widehat{\boldsymbol{\alpha}}_j^T \partial_k \mathbf{K}_n(\mathbf{x}_{i\widehat{\mathcal{S}}_t})\right\}^2. \tag{3.3}$$

Note that $\widehat{\boldsymbol{\alpha}}_j$ is obtained in (3.2), and thus (3.3) can be directly computed in a parallel fashion since $\partial_k \mathbf{K}_n(\mathbf{x}_{i\widehat{\mathcal{S}}_t})$ is known once $K(\cdot,\cdot)$ is specified. Then, the estimated directed edges can be denoted as $\widehat{\mathcal{E}}_j = \big\{k \to j, \|\widehat{g}_{jk}\|_n^2 > v_n^{(t)}, \text{for any } k \in \widehat{\mathcal{S}}_t\big\}$ for some pre-specified $v_n^{(t)}$.

Algorithm 1 The proposed algorithm

1: Input: sample matrix $\mathbf{X} \in \mathcal{R}^{n\times p}$, $\widehat{\mathcal{S}} = \emptyset$, and $t = 0$;
2: Until $\widehat{\mathcal{S}} = \mathcal{V}$:

a. For any $j \in \mathcal{V}\backslash\{\widehat{\mathcal{S}}\}$, compute the conditional variance $\widehat{E\mathrm{Var}}(x_j|\mathbf{x}_{\widehat{\mathcal{S}}})$;

b. Define $\widehat{\mathcal{A}}_t = \big\{k, |\widehat{E\mathrm{Var}}(x_k|\mathbf{x}_{\widehat{\mathcal{S}}}) - \widehat{\sigma}^{(t)}_{\min}| < \epsilon_t\big\}$;

c. Define $\widehat{\mathcal{E}}_j = \big\{k \to j, \|\widehat{g}_{jk}\|_n^2 > v_n^{(t)}, \text{for any } k \in \widehat{\mathcal{S}}\big\}$ for any $j \in \widehat{\mathcal{A}}_t$;

d. Let $\widehat{\mathcal{S}} = \widehat{\mathcal{S}} \cup \widehat{\mathcal{A}}_t$;

e. $t \leftarrow t + 1$;

3: Let $\widehat{T} = t$.
4: Return: $\{\widehat{\mathcal{A}}_t\}_{t=0}^{\widehat{T}-1}$ and $\{\widehat{\mathcal{E}}_j\}_{j\in\mathcal{V}}$.

We repeat the above reconstructing procedure until all the nodes have been assigned and all the directed relations have been recovered.

It is thus clear that the proposed method is motivated by our identifiability result in Theorem 1, which takes the advantage of topological layer to assure acyclicity and facilitate DAG learning, and kernel ridge regressions are used as efficient tools to reconstruct layers and recover parent-child relations. This significantly differs from the learning sparse nonparametric DAG method (NOTEARS, Zhang et al., 2020) from a methodological point of view. Specifically, NOTEARS is a score-based method that it searches over the space consisting of all the possible graphs, and a gradient-based criteria is developed to force the graph to be acyclic. Then, some pre-specified modeling is used to evaluate the score function, including linear model, additive model or neural network, and finally, the graph minimizing the score is returned. Theoretically, the asymptotic properties of the proposed method are established in terms of exact DAG recovery under mild conditions in Section 4, yet the theoretical properties of NOTEARS remain largely unknown.

3.2. Tuning

Note that the numerical performance of the proposed method depends on the choice of tuning parameters ϵ_t and $v_n^{(t)}$. For selecting the optimal values of $\{v_n^{(t)}\}'s$, we follow the suggestion of He, Wang and Lv (2021). For the parameters ϵ_t, we employ the stability-based criterion (Sun, Wang and Fang, 2013) to select the optimal value, which is also used in Zhao, He and Wang (2022). The key idea is to measure the stability of topological layer reconstruction by randomly splitting the training sample into two parts and comparing the disagreement between the two estimated active sets. Specifically, given a value ϵ, we randomly split the training sample $\mathcal{Z}^M$ into two parts $\mathcal{Z}_1^M$ and $\mathcal{Z}_2^M$. Then the proposed

method is applied to $\mathcal{Z}_1^M$ and $\mathcal{Z}_2^M$ and obtains two estimated active sets $\widehat{\mathcal{A}}_{1,\epsilon}$ and $\widehat{\mathcal{A}}_{2,\epsilon}$, respectively. The disagreement between $\widehat{\mathcal{A}}_{1,\epsilon}$ and $\widehat{\mathcal{A}}_{2,\epsilon}$ is measured by Cohen's kappa coefficient

$$\kappa(\widehat{\mathcal{A}}_{1,\epsilon}, \widehat{\mathcal{A}}_{2,\epsilon}) = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where $Pr(a) = (n_{11} + n_{22})/p$ and $Pr(e) = \{(n_{11} + n_{12})(n_{11} + n_{21})\}/p^2 + \{(n_{12} + n_{22})(n_{21} + n_{22})\}/p^2$ with $n_{11} = |\widehat{\mathcal{A}}_{1,\epsilon} \cap \widehat{\mathcal{A}}_{2,\epsilon}|, n_{12} = |\widehat{\mathcal{A}}_{1,\epsilon} \cap \widehat{\mathcal{A}}_{2,\epsilon}^C|, n_{21} = |\widehat{\mathcal{A}}_{1,\epsilon}^C \cap \widehat{\mathcal{A}}_{2,\epsilon}|, n_{22} = |\widehat{\mathcal{A}}_{1,\epsilon}^C \cap \widehat{\mathcal{A}}_{2,\epsilon}^C|$ and $|\cdot|$ denotes the set cardinality. The procedure is repeated for B times and the topological layer reconstruction stability is measured as

$$\hat{s}(\Psi_\epsilon) = \frac{1}{B} \sum_{b=1}^{B} \kappa(\widehat{\mathcal{A}}_{1,\epsilon}^b, \widehat{\mathcal{A}}_{2,\epsilon}^b).$$

Finally, the selected parameter ϵ is set as $\max\left\{\epsilon : \hat{s}(\Psi_\epsilon)/\max_\epsilon \hat{s}(\Psi_\epsilon) \geq d\right\}$, where $d \in (0, 1)$ is some given percentage. Note that the adopted selection criteria (Sun, Wang and Fang, 2013) is originally designed for the purpose of variable selection with theoretical guarantees, and the ratio is used to avoid missing some weak signals. Moreover, the choice of maximum can be regarded as a pre-specified parameter and in practice, one can also use minimum, mean or median.

4. Statistical Guarantees

In this section, we investigate the theoretical property of the proposed method in terms of exact DAG recovery. The asymptotic theoretical results are established by using the kernel ridge regression and learning gradients in the smooth RKHS under some regularity assumptions. For theoretical analysis, we define some intermediate target functions and introduce some functional operators. Specifically, for any $t = 1, \ldots, T-1$ and $j \in \mathcal{V} \backslash \{\mathcal{S}_t\}$, we define $f_{j,\mathcal{S}_t}^*(\mathbf{x}_{\mathcal{S}_t}) = \operatorname{argmin}_f E\{x_j - f(\mathbf{x}_{\mathcal{S}_t})\}^2$ and it is clear that $f_{j,\mathcal{S}_t}^*(\mathbf{x}_{\mathcal{S}_t}) = E(x_j|\mathbf{x}_{\mathcal{S}_t})$. We further assume that $f_{j,\mathcal{S}_t}^*(\mathbf{x}_{\mathcal{S}_t}) \in \mathcal{H}_K$ and it is worth noting that $E(x_j|\mathbf{x}_{\mathcal{S}_t}) = E(x_j|\mathbf{x}_{\mathrm{pa}_j})$ if $j \in \mathcal{A}_t$ and $E(x_j|\mathbf{x}_{\mathcal{S}_t}) \neq E(x_j|\mathbf{x}_{\mathrm{pa}_j})$ if $j \in \mathcal{V} \backslash \{\mathcal{S}_t \cup \mathcal{A}_t\}$. We denote the supports of $\mathbf{x}_{\mathcal{S}_t}$ as $\mathcal{X}_t \subset \mathcal{X}$, which are assumed to be compact. Without loss of generality, we also assume that the K-norms of all the target functions $f_{j,\mathcal{S}_t}^*$ are upper bounded by $R/2$ for mathematical simplicity throughout this paper, where R denotes some positive quantity, and this technical requirement can be easily satisfied by taking R relatively large. Note that the compactness condition is commonly assumed in machine learning literature (Smale and Zhou, 2007; Rosasco et al., 2013; Lv et al., 2018) to ensure universality and the Mercer's theorem, which also implies that all the noise terms $\{n_j\}_{j \in \mathcal{V}}$ have compact support and recently, many efforts have been made to extend it to the non-compact setting (Steinwart and Scovel, 2011; Simon-Gabriel and Schölkopf, 2018). Moreover, we introduce the integral

operators $L_{K,t} : \mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}}) \to \mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})$ that

$$L_{K,t}(f)(\mathbf{x}_{\mathcal{S}_t}) = \int K(\mathbf{x}_{\mathcal{S}_t}, \mathbf{u}_{\mathcal{S}_t}) f(\mathbf{u}_{\mathcal{S}_t}) d\rho_{\mathbf{x}_{\mathcal{S}_t}}(\mathbf{u}_{\mathcal{S}_t}),$$

for any $f \in \mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}}) = \{f : \int f^2(\mathbf{x}_{\mathcal{S}_t}) d\rho_{\mathbf{x}_{\mathcal{S}_t}} < \infty\}$.

We first establish the layer recovery consistency based on kernel ridge regression. The following technical assumptions are required to establish consistency.

Assumption 2. *For any $t = 1, \ldots, T-1$ and $j \in \mathcal{V} \backslash \{\mathcal{S}_t\}$, suppose that $f^*_{j,\mathcal{S}_t}$ is in the range of the r-th power of $L_{K,t}$, denoted as $L^r_{K,t}$, for some positive constant $r \in (1/2, 1]$.*

Assumption 3. *There exist some constants κ_1 and κ_2 such that for any $\mathcal{S} \subset \mathcal{V}$, there hold $\sup \|K_{\mathbf{x}_\mathcal{S}}\|_K \leq \kappa_1$ and $\sup \|\partial_k K_{\mathbf{x}_\mathcal{S}}\|_K \leq \kappa_2$.*

Note that the fractional operators $L^r_{K,t}$ in Assumption 2 make sense as the operator $L_{K,t}$ on $\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})$ is self-adjoint and semi-positive definite. As pointed out by Smale and Zhou (2007), the requirement that $r \geq 1/2$ is a general assumption, which ensures that the range of $L^r_{K,t}$ is contained in $\mathcal{H}_K$ (Smale and Zhou, 2007), and thus we can deduce that there exists some function $h_{j,t} \in \mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})$ such that $f^*_{j,\mathcal{S}_t} = L^r_{K,t} h_{j,t} \in \mathcal{H}_K$. This ensures strong estimation consistency under the RKHS-norm. Assumption 3 requires the kernel function and its gradient function to be upper bounded, which is commonly assumed in machine learning literature (Rosasco et al., 2013) and is satisfied by many kernel functions, including the Gaussian kernel.

Theorem 3. *Suppose that Assumptions 1 to 3 are satisfied. Then, for any $\zeta > 0$ and $k \in \mathcal{V}$, we have*

$$P\Big(|\widehat{\mathrm{Var}}(x_k) - \mathrm{Var}(x_k)| > \zeta\Big) \leq 2 \exp\bigg(-\frac{n\zeta^2}{2C^4_{\mathcal{X}}}\bigg),$$

where $C_{\mathcal{X}}$ denotes the diameter of the support $\mathcal{X}$. Additionally, if we take $\epsilon_0 = M_{\max}/2$, there holds

$$P\big(\widehat{\mathcal{A}}_0 = \mathcal{A}_0\big) \geq 1 - 2p \exp\bigg(-\frac{nM^2_{\max}}{32C^4_{\mathcal{X}}}\bigg).$$

Theorem 3 establishes the estimation consistency of the variance estimator and ensures that the first layer $\mathcal{A}_0$ can be exactly reconstructed with high probability. It is worth pointing out that the consistency result still holds if we take $\epsilon_0 \in (C_1\sqrt{\log(2p)/n}, M_{\max}/2]$ for some positive constant C_1. Once $\mathcal{A}_0$ has been reconstructed, the subsequent layers can also be reconstructed in a sequence.

To establish the asymptotic results for the lower layers, we define the event that

$$\mathcal{J} = \bigcap_t \Big\{ \max_{j \in \mathcal{V} \setminus \{\mathcal{S}_t\}} \|\widehat{f}_j\|_K \le R \Big\},$$

and use $\mathcal{J}^c$ to denote its complementary. Without loss of generality, we assume that the layers $\mathcal{A}_1, \ldots, \mathcal{A}_{t-1}$ have been exactly reconstructed and the following theorem ensures that the layer $\mathcal{A}_t$ can be recovered with high probability by using kernel-based estimation under mild conditions.

Theorem 4. *Suppose that all the assumptions in Theorem* 3 *are satisfied. Given the events* $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\}$, $t \ge 1$ *and the event* $\mathcal{J}$, *if we take* $\lambda = n^{-1/(2r+1)}$, *then for any* $\zeta > 0$ *and* $j \in \mathcal{V} \setminus \{\mathcal{S}_t\}$, *there holds*

$$\begin{aligned} &P\Big(\big| \widehat{E}\{\widehat{\mathrm{Var}}(x_j | \mathbf{x}_{\mathcal{S}_t})\} - E\{\mathrm{Var}(x_j | \mathbf{x}_{\mathcal{S}_t})\} \big| > \zeta \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J} \Big) \\ &\le 2 \exp\Big(- \frac{8n\zeta^2}{C_{\mathcal{X}}^4} \Big) + 4 \exp\Big(- \frac{n^{(2r-1)/\{2(2r+1)\}} \zeta}{2C_{jt}} \Big), \end{aligned}$$

where $C_{jt} = 6\kappa_1^2 R \max\{2\kappa_1 \max\{C_{\mathcal{X}} + 2\kappa_1 R, \sqrt{2(2\kappa_1^2 R^2 + \sigma_j^2)}\}, \sqrt{2}, \|L_{K,t}^{-r} f_{j,\mathcal{S}_t}^*\|_2\}$.

Additionally, if we take $\epsilon_t = M_{\max}/2$, *there holds*

$$\begin{aligned} &P(\widehat{\mathcal{A}}_t = \mathcal{A}_t \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) \\ &\ge 1 - 2(p - |\mathcal{S}_t|) \exp\Big(- \frac{nM_{\max}^2}{2C_{\mathcal{X}}^4} \Big) - 4(p - |\mathcal{S}_t|) \exp\Big(- \frac{M_{\max} n^{(2r-1)/\{2(2r+1)\}}}{8C_{jt}} \Big). \end{aligned}$$

The first part of Theorem 4 shows that the estimated criteria converges to the truth with high probability, which plays a crucial role to establish the layer recovery consistency. The second part of Theorem 4 ensures that the layer $\mathcal{A}_t$ can be exactly reconstructed under mild conditions with some proper choice of ϵ_t. In fact, the consistency result still holds if we take ϵ_t in an interval with upper bound $M_{\max}/2$ following the similar choice for ϵ_0. The proof of Theorem 4 is completed by using Lemma S1 in the supplementary file. As a direct consequence of Theorems 3 and 4, all the layers can be exactly reconstructed with high probability.

We want to emphasize that once the layer $\mathcal{A}_t$ has been constructed, the parent-child relations between nodes in $\mathcal{A}_t$ and $\mathcal{S}_t = \cup_{d=0}^{t-1} \mathcal{A}_d$ can be obtained directly by computing the estimated gradient function using (3.3) without any extra estimation. More importantly, the selection consistency of the parent set for nodes in $\mathcal{A}_t$ can also be established under mild conditions. The following technical assumption is needed to establish the recovery consistency of parent-child relations.

Assumption 4. *For any* $t = 1, \ldots, T-1$ *and* $j \in \mathcal{A}_t$, *there exists some positive constant* C_2 *such that* $\min_{k \in pa_j} \|g_{jk}^*\|_2^2 > C_2 n^{-(2r-1)/\{2(2r+1)\}}$

$\{\log(4|\mathcal{S}_t|\max\{n,|\mathcal{S}_t|\})\}^{\beta}$ *for some* $\beta > 1$ *and* $\max_{k\in\mathcal{S}_t\setminus\{pa_j\}} \|g^*_{jk}\|_2^2 = 0$.

Note that by the definition of $\mathcal{A}_t$, there holds $\mathrm{pa}_j \subset \mathcal{S}_t$ for any $j \in \mathcal{A}_t$, and Assumption 4 is a general condition that requires all the parents should make a contribution to their child by assuming that given all the other nodes belonging to pa_j, the true gradient function contains sufficient information about parent nodes. This is equivalent to assuming each true function in (2.1) should not be a constant in any of its arguments and is also known as the causal minimality condition in DAG learning literature (Peters et al., 2014).

Lemma 1. *Suppose that all the assumptions in Theorem* 4 *as well as Assumption* 4 *are satisfied. Then, for any* $t \geq 1$, *given the events that* $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_t = \mathcal{A}_t\}$ *and the event* $\mathcal{J}$, *and if we take* $v_n^{(t)} = (C_2/2)n^{-2r-1/\{2(2r+1)\}}\{\log(4|\mathcal{S}_t|\max\{n,|\mathcal{S}_t|\})\}^{\beta}$, *there holds*

$$P\big(\{\mathcal{E}_j = \widehat{\mathcal{E}}_j : j \in \widehat{\mathcal{A}}_t\}|\widehat{\mathcal{A}}_0 = \mathcal{A}_0, \ldots, \widehat{\mathcal{A}}_t = \mathcal{A}_t, \mathcal{J}\big) \geq 1 - \frac{1}{\max\{n,|\mathcal{S}_t|\}}.$$

Lemma 1 shows that the parent set for nodes in $\mathcal{A}_t$ can be also consistently recovered after the layer $\mathcal{A}_t$ is correctly reconstructed. It is interesting to point out that Lemma 1 is particularly attractive in that it is established without any further estimation after reconstructing $\mathcal{A}_t$, due to the fact that the gradient functions can be directly computed as a by-product as illustrated by (3.3). Now, we turn to establish the exact DAG recovery consistency of the proposed method.

Theorem 5. *Suppose that all the assumptions of Lemma* 1 *are satisfied. Then, we have*

$$P(\widehat{\mathcal{G}} = \mathcal{G}) \to 1, \ \text{as } n \to \infty.$$

Theorem 5 ensures that the DAG $\mathcal{G}$ can be consistently recovered by the proposed method with probability tending to 1. Note that the proof of Theorem 5 is conducted by using the fact that $P(\widehat{\mathcal{G}} \neq \mathcal{G}) \leq P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}) + P(\mathcal{J}^c)$, and directly by the results in Smale and Zhou (2007), $P(\mathcal{J}^c) \to 0$ as $n \to \infty$ under some mild conditions. It is particularly attractive in the literature on DAG learning in the sense that it allows for general parent-child relations and it provides a solid theoretical guarantee for learning nonparametric DAG in terms of exact DAG recovery.

5. Numerical Experiments

In this section, we compare the numerical performance of the proposed method by using centralized Gaussian kernel (He et al., 2022), denoted as NL, against some state-of-the-art methods, including the nonparametric variance-based algorithm with additive modeling (NPVAR; Gao, Ding and Aragam, 2020), the nonparametric regression with independence test (RESIT; Peters et al.,

2014), the nonparametric additive method (CAM; Bühlmann, Peters and Ernest, 2013), the greedy equivalent search algorithm (GES; Chickering, 2003), the high-dimensional constraint-based PC algorithm (PC; Kalisch and Bühlmann, 2007) and NOTEARS (Zhang et al., 2020). We code NL in `R` and implement CAM by using the `R` package `CAM`. Both methods PC and GES are implemented by using the `R` package `pcalg`. The `R` codes of NPVAR and RESIT are available online at `http://people.tuebingen.mpg.de/jpeters/onlineCodeANM.zip` and `https://github.com/MingGao97/NPVAR`, respectively. The `Python` code of NOTEARS is available at `https://github.com/xunzheng/notears`. Note that NPVAR, RESIT and CAM fit the nonparametric functions under the additive modeling, and thus their performance highly relies on the validity of the additive model assumption.

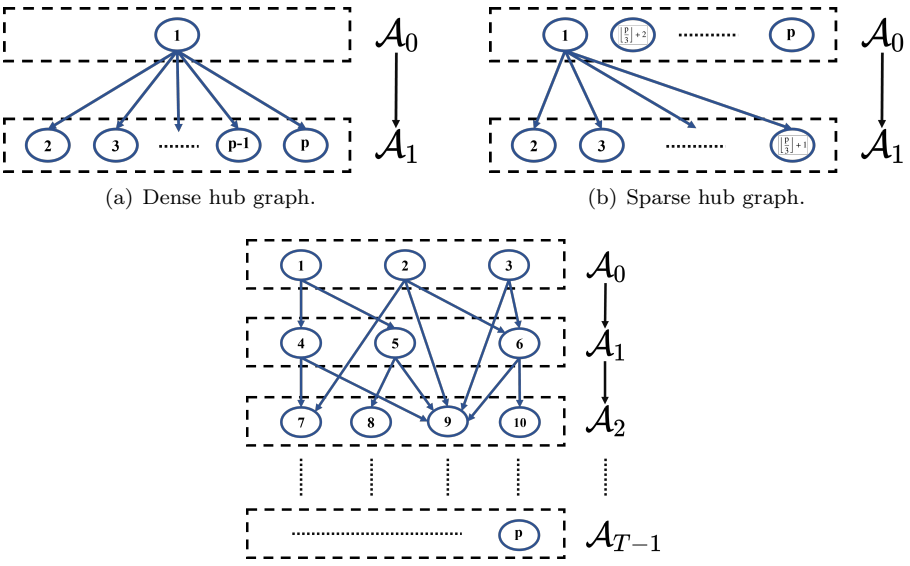
To evaluate the performance of all the methods, we report the true positive rate (TPR) and false discovery rate (FPR) to evaluate the accuracy of estimated directed edges. We also employ the normalized structural Hamming distance (HD; Tsamardinos, Brown and Aliferis, 2006) to evaluate the closeness of the true and estimated DAG, and use the Matthews correlation coefficient (MCC; Yuan et al., 2019) to overall accuracy of the estimated DAG structure. Note that the metric HD measures the smallest number of edge insertions, deletions, and flips to convert the estimated DAG into the truth DAG. It is worth noting that small values of HD, FDR and FPR, but large values of TPR and MCC indicate a good reconstruction of a DAG.

5.1. Simulated examples

In this section, we examine the numerical performance of all the competitors in three simulated examples, where Examples 1 and 2 consider a dense and sparse hub graph, respectively, and Example 3 considers a random graph generated by the Erdös Rényi (ER) model.

Example 1. we consider a DAG where the only directed structure is edged directing from the first node, known as the hub node, to all the other nodes. Clearly, we have $T = 2, \mathcal{A}_0 = \{1\}$ and $\mathcal{A}_1 = \{2, \ldots, p\}$. Example 1 is illustrated in Figure 2(a). Specifically, we generate $n_1 \sim U(-0.5, 0.5)$ and $x_j, j \in \mathcal{A}_1$, from $x_j = f_j^*(x_1) + n_j$, where $n_j \sim U(-1, 1)$ and $f_j^*(x)$ is randomly chosen from $f^{(1)}(x) = 0.3\sin(\pi x) + 0.3\cos(\pi x) + 0.4\sin^2(\pi x)$, $f^{(2)}(x) = 0.2\cos^3(\pi x) + 0.2\sin^3(\pi x)$, $f^{(3)}(x) = \arctan(x)$, $f^{(4)}(x) = \sin(\pi x)/\{2 - \sin(\pi x)\}$ with equal probability and is also centered.

Example 2. The generated DAG is the same as that in Example 1 except that the first node is only directed to the next $\lfloor p/3 \rfloor + 1$ nodes and all the remaining nodes are isolated. Clearly, we have $T = 2, \mathcal{A}_0 = \{1, \lfloor p/3 \rfloor + 2, \ldots, p\}$ and $\mathcal{A}_1 = \{2, \ldots, \lfloor p3 \rfloor + 1\}$, and the structure of the underlying DAG is illustrated in Figure 2(b).



(a) Dense hub graph. (b) Sparse hub graph.

(c) Random graph.

Figure 2. The topological layer of the DAG structures in Examples 1 to 3.

Example 3. We consider a random DAG generated by the ER graph and the underlying structure is illustrated in Figure 2(c). The probability of connecting an edge is set as $P_E = 0.25$ for $p = 5$ and 20, and $P_E = 0.05$ for $p = 100$. Specifically, we generate $n_j \sim U(-0.5, 0.5)$ for any $j \in \mathcal{A}_0$, $n_j \sim U(-1.5, 1.5)$ for any $j \in \mathcal{A}_1$ and $\mathcal{A}_2$, and $n_j \sim U(-3, 3)$ for any $\in \mathcal{A}_3, \ldots, \mathcal{A}_{T-1}$, and set $T = \max\{4, \lfloor \log(p) \rfloor\}$. For $t \geq 1$, $x_j, j \in \mathcal{A}_t$, is generated by $x_j = \sum_{0 \leq s \leq t-1} \sum_{k \in \mathcal{A}_s \cap \mathrm{pa}_j} \theta_{sk} f^{(s)}(x_k) + \sum_{k,\ell \in \mathrm{pa}_j, k \neq \ell} \{\prod_{1 \leq s \leq t} (1 - \theta_{sk})(1 - \theta_{s\ell})\} g_1(x_k) g_2(x_\ell) + n_j$ with $\theta_{sk} \sim \mathrm{Bern}(1, 0.75)$. For the main function, we set $f^{(0)}(x) = 0.2\cos^3(\pi x) + 0.2\sin^3(\pi x)$, $f^{(1)}(x) = 1.5\arctan(x)$, $f^{(2)}(x) = 3\sqrt{|x|}$, $f^{(3)}(x) = 0.3\sin(\pi x) + 0.3\cos(\pi x) + 0.4\sin^2(\pi x)$ and $f^{(t)}(x) = \sin(\pi x)/\{2 - \sin(\pi x)\}, t \geq 4$, and consider $g_1(x) = 2|x|^{0.4}$ and $g_2(x) = \exp\{\sin(\pi x)\}$ for the interaction term. Note that the parameters θ_{sk} are used to ensure the parent nodes in the main function and interaction term are distinct, for example, if $5 \in \mathcal{A}_2$ and $\mathrm{pa}_5 = \{1, 2, 3, 4\}$ with $\{1, 3\} \subset \mathcal{A}_0$, $\{2, 4\} \subset \mathcal{A}_1$, then a possible generating scheme of x_5 is $x_5 = f^{(0)}(x_1) + f^{(1)}(x_2) + g_1(x_3) g_2(x_4) + n_5$ and the functions are also centered.

For each example, we repeat the data generating scheme 50 times and the averaged performance of all the competitors under the cases by varying n and p from $\{100, 200, 500\}$ and $\{5, 20, 100\}$, respectively, are provided in Tables 1 to 3. Note that $**$ is used to denote the fact that the corresponding methods take too long to produce any results or is not applicable.

It is evident from Tables 1 to 3 that NL outperforms all the other competitors in almost all the cases, except in Example 3 where NL is the second performer

Table 1. The averaged performance metrics of different methods as well as their standard errors in parentheses in Example 1. Here, ** denotes that the corresponding methods took too long to produce any results or was not applicable.

p	n	Methods	HD	FDR	TPR	MCC
5	100	NL	0.0520(0.0101)	0.0000(0.0000)	0.7400(0.0505)	0.8159(0.0358)
		NPVAR	0.1380(0.0055)	0.1400(0.0496)	0.3100(0.0273)	0.4790(0.0321)
		CAM	0.2110(0.0096)	0.5533(0.0509)	0.2500(0.0295)	0.2431(0.0381)
		RESIT	0.2160(0.0123)	0.5353(0.0538)	0.2650(0.0298)	0.2573(0.0414)
		GES	0.2790(0.0163)	0.7103(0.0463)	0.2550(0.0388)	0.1401(0.0492)
		PC	0.1290(0.0111)	0.2800(0.0502)	0.4850(0.0394)	0.5363(0.0469)
		NOTEARS	0.1920(0.0056)	0.6544(0.0466)	0.2750(0.0399)	0.2502(0.0350)
	200	NL	0.0370(0.0090)	0.0000(0.0000)	0.8150(0.0451)	0.8699(0.0318)
		NPVAR	0.0850(0.0066)	0.0200(0.0200)	0.5750(0.0329)	0.7116(0.0267)
		CAM	0.2380(0.0107)	0.6113(0.0338)	0.3050(0.0279)	0.2322(0.0336)
		RESIT	0.1790(0.0113)	0.3727(0.0365)	0.4800(0.0235)	0.4583(0.0300)
		GES	0.2910(0.0162)	0.7470(0.0428)	0.2450(0.0413)	0.1114(0.0494)
		PC	0.0980(0.0105)	0.1900(0.0332)	0.6550(0.0342)	0.6816(0.0362)
		NOTEARS	0.1800(0.0047)	0.5667(0.0540)	0.2750(0.0359)	0.2889(0.0350)
20	200	NL	0.0000(0.0000)	0.0000(0.0000)	0.9989(0.0011)	0.9994(0.0006)
		NPVAR	0.0378(0.0005)	0.0147(0.0071)	0.2484(0.0095)	0.4809(0.0099)
		CAM	0.0428(0.0010)	0.3911(0.0142)	0.3968(0.0141)	0.4701(0.0140)
		RESIT	0.1109(0.0020)	0.9339(0.0064)	0.0884(0.0083)	0.0213(0.0077)
		GES	0.0925(0.0010)	0.8715(0.0059)	0.1453(0.0065)	0.0903(0.0064)
		PC	0.0483(0.0015)	0.4509(0.0316)	0.2495(0.0138)	0.3490(0.0212)
		NOTEARS	0.0656(0.0013)	0.6709(0.0134)	0.2874(0.0133)	0.2717(0.0110)
	500	NL	0.0000(0.0000)	0.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
		NPVAR	0.0210(0.0008)	0.0040(0.0029)	0.5832(0.0158)	0.7507(0.0104)
		CAM	0.0335(0.0011)	0.3165(0.0120)	0.6168(0.0127)	0.6320(0.0120)
		RESIT	0.1483(0.0025)	0.9192(0.0031)	0.1853(0.0062)	0.0538(0.0047)
		GES	0.0934(0.0008)	0.8793(0.0047)	0.1368(0.0050)	0.0818(0.0051)
		PC	0.0480(0.0028)	0.4690(0.0398)	0.3740(0.0283)	0.4220(0.0346)
		NOTEARS	0.0553(0.0010)	0.5568(0.0193)	0.2726(0.0121)	0.3142(0.0094)
100	200	NL	0.0000(0.0000)	0.0000(0.0000)	0.9990(0.0004)	0.9995(0.0002)
		NPVAR	**	**	**	**
		CAM	**	**	**	**
		RESIT	0.0340(0.0006)	0.9956(0.0008)	0.0103(0.0019)	-0.0088(0.0013)
		GES	0.0334(0.0002)	0.9838(0.0007)	0.0388(0.0017)	0.0098(0.0011)
		PC	**	**	**	**
		NOTEARS	0.0153(0.0001)	0.9936(0.0036)	0.0036(0.0021)	-0.0024(0.0028)
	500	NL	0.0000(0.0000)	0.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
		NPVAR	**	**	**	**
		CAM	**	**	**	**
		RESIT	0.0551(0.0008)	0.9917(0.0003)	0.0378(0.0015)	-0.0036(0.0008)
		GES	0.0297(0.0001)	0.9826(0.0008)	0.0354(0.0016)	0.0108(0.0012)
		PC	**	**	**	**
		NOTEARS	0.0118(0.0001)	0.9420(0.0152)	0.0162(0.0042)	0.0262(0.0080)

Table 2. The averaged performance metrics of different methods as well as their standard errors in parentheses in Example 2. Here, ** denotes that the corresponding methods took too long to produce any results or was not applicable.

p	n	Methods	HD	FDR	TPR	MCC
5	100	NL	0.1150(0.0056)	0.4177(0.0149)	0.9600(0.0154)	0.7027(0.0140)
		NPVAR	0.1090(0.0053)	0.3600(0.0686)	0.2733(0.0352)	0.3971(0.0448)
		CAM	0.1700(0.0074)	0.6850(0.0509)	0.2133(0.0340)	0.1887(0.0411)
		RESIT	0.1530(0.0091)	0.5133(0.0545)	0.3000(0.0319)	0.3166(0.0413)
		GES	0.2120(0.0137)	0.7317(0.0494)	0.2600(0.0460)	0.1685(0.0527)
		PC	0.1120(0.0117)	0.4230(0.0606)	0.4600(0.0529)	0.4680(0.0603)
		NOTEARS	0.1390(0.0036)	0.7200(0.0508)	0.2333(0.0439)	0.2247(0.0394)
	200	NL	0.0640(0.0061)	0.2590(0.0230)	0.9667(0.0143)	0.8163(0.0172)
		NPVAR	0.0610(0.0050)	0.0400(0.0280)	0.5933(0.0334)	0.7269(0.0289)
		CAM	0.1740(0.0086)	0.6033(0.0414)	0.3267(0.0323)	0.2802(0.0382)
		RESIT	0.1200(0.0100)	0.3130(0.0414)	0.5400(0.0342)	0.5427(0.0356)
		GES	0.2480(0.0138)	0.8300(0.0443)	0.1733(0.0449)	0.0593(0.0506)
		PC	0.0770(0.0103)	0.2400(0.0460)	0.6600(0.0431)	0.6730(0.0473)
		NOTEARS	0.1350(0.0050)	0.7257(0.0550)	0.2200(0.0472)	0.2184(0.0444)
20	200	NL	0.0086(0.0007)	0.1304(0.0127)	0.8923(0.01204)	0.8751(0.0100)
		NPVAR	0.0249(0.0005)	0.0157(0.0091)	0.2785(0.0132)	0.5090(0.0132)
		CAM	0.0303(0.0008)	0.4153(0.0179)	0.4092(0.0155)	0.4729(0.0158)
		RESIT	0.0669(0.0014)	0.8893(0.0096)	0.1277(0.0105)	0.0856(0.0103)
		GES	0.0633(0.0009)	0.8452(0.0074)	0.1877(0.0088)	0.1392(0.0082)
		PC	0.0295(0.0012)	0.3775(0.0281)	0.3662(0.0181)	0.4626(0.0223)
		NOTEARS	0.0389(0.0008)	0.5870(0.0177)	0.2892(0.0170)	0.3198(0.0137)
	500	NL	0.0047(0.0005)	0.0000(0.0000)	0.8631(0.0138)	0.9255(0.0077)
		NPVAR	0.0119(0.0005)	0.0150(0.0060)	0.6615(0.0144)	0.8001(0.0097)
		CAM	0.0271(0.0009)	0.3878(0.0142)	0.5985(0.0132)	0.5907(0.0127)
		RESIT	0.0878(0.0020)	0.8768(0.0061)	0.2446(0.0098)	0.1330(0.0080)
		GES	0.0632(0.0007)	0.8464(0.0058)	0.1862(0.0073)	0.1380(0.0066)
		PC	0.0275(0.0017)	0.3828(0.0292)	0.5323(0.0251)	0.5587(0.0273)
		NOTEARS	0.0351(0.0007)	0.5016(0.0245)	0.2815(0.0143)	0.3496(0.0137)
100	200	NL	0.0020(0.0000)	0.0016(0.0008)	0.6963(0.0069)	0.8324(0.0041)
		NPVAR	**	**	**	**
		CAM	**	**	**	**
		RESIT	0.0211(0.0003)	0.9928(0.0013)	0.0155(0.0028)	0.0008(0.0019)
		GES	0.0266(0.0002)	0.9833(0.0007)	0.0515(0.0022)	0.0179(0.0013)
		PC	**	**	**	**
		NOTEARS	0.0103(0.0002)	0.9487(0.0101)	0.0294(0.0059)	0.0334(0.0075)
	500	NL	0.0018(0.0000)	0.0000(0.0000)	0.7333(0.0075)	0.8550(0.0044)
		NPVAR	**	**	**	**
		CAM	**	**	**	**
		RESIT	0.0329(0.0006)	0.9861(0.0007)	0.0548(0.0024)	0.0146(0.0014)
		GES	0.0219(0.0001)	0.9791(0.0010)	0.0497(0.0022)	0.0223(0.0015)
		PC	**	**	**	**
		NOTEARS	0.0077(0.0001)	0.7380(0.0169)	0.0803(0.0052)	0.1410(0.0090)

Table 3. The averaged performance metrics of different methods as well as their standard errors in parentheses in Example 3. Here, ** denotes that the corresponding methods took too long to produce any results or was not applicable.

p	n	Methods	HD	FDR	TPR	MCC
5	100	NL	0.1530(0.0123)	0.4274(0.0331)	0.9753(0.0108)	0.6797(0.0255)
		NPVAR	0.1260(0.0082)	0.2200(0.0592)	0.3168(0.0314)	0.4629(0.0387)
		CAM	0.1670(0.0116)	0.4797(0.0619)	0.2456(0.0324)	0.3005(0.0427)
		RESIT	0.1640(0.0132)	0.3343(0.0602)	0.2687(0.0267)	0.3637(0.0403)
		GES	0.2220(0.0154)	0.6717(0.0484)	0.1822(0.0252)	0.1514(0.0360)
		PC	0.2000(0.0141)	0.6300(0.0614)	0.1520(0.0257)	0.1680(0.0412)
		NOTEARS	0.1430(0.0107)	0.3629(0.0570)	0.3591(0.0347)	0.4226(0.0402)
	200	NL	0.1120(0.0108)	0.3520(0.0340)	0.9603(0.0134)	0.7341(0.0253)
		NPVAR	0.1230(0.0075)	0.2600(0.0627)	0.3209(0.0339)	0.4541(0.0417)
		CAM	0.1720(0.0106)	0.5433(0.0613)	0.2279(0.0308)	0.2631(0.0421)
		RESIT	0.1530(0.0141)	0.3200(0.0579)	0.3532(0.0394)	0.4236(0.0452)
		GES	0.2240(0.0150)	0.6637(0.0538)	0.1977(0.0325)	0.1608(0.0424)
		PC	0.2060(0.0143)	0.6330(0.0591)	0.1690(0.0296)	0.1700(0.0420)
		NOTEARS	0.1400(0.0081)	0.3300(0.0613)	0.2989(0.0313)	0.4030(0.0380)
20	200	NL	0.1159(0.0034)	0.5246(0.0117)	0.7376(0.0116)	0.5331(0.0100)
		NPVAR	0.0715(0.0020)	0.0045(0.0027)	0.3236(0.0123)	0.5425(0.0104)
		CAM	0.0800(0.0035)	0.2469(0.0245)	0.3713(0.0157)	0.4921(0.0192)
		RESIT	0.1281(0.0045)	0.6861(0.0227)	0.1296(0.0074)	0.1408(0.0136)
		GES	0.1507(0.0044)	0.8414(0.0115)	0.0981(0.0080)	0.0493(0.0105)
		PC	0.1200(0.0038)	0.6970(0.0272)	0.0920(0.0084)	0.1170(0.0159)
		NOTEARS	0.1227(0.0050)	0.5379(0.0147)	0.5141(0.0161)	0.4163(0.0102)
	500	NL	0.0551(0.0019)	0.2241(0.0124)	0.6874(0.0109)	0.6993(0.0078)
		NPVAR	0.0538(0.0016)	0.0093(0.0048)	0.4914(0.0144)	0.6746(0.0109)
		CAM	0.0781(0.0037)	0.2853(0.0187)	0.4682(0.0138)	0.5384(0.0153)
		RESIT	0.1381(0.0059)	0.6715(0.0194)	0.2228(0.0121)	0.1966(0.0156)
		GES	0.1576(0.0045)	0.8294(0.0103)	0.1250(0.0079)	0.0648(0.0092)
		PC	0.1250(0.0036)	0.7090(0.0169)	0.1250(0.0079)	0.1340(0.0121)
		NOTEARS	0.0876(0.0036)	0.3227(0.0224)	0.4358(0.0147)	0.4905(0.0118)
100	200	NL	0.0600(0.0018)	0.8098(0.0062)	0.4856(0.0064)	0.2769(0.0051)
		NPVAR	**	**	**	**
		CAM	**	**	**	**
		RESIT	0.0339(0.0006)	0.9058(0.0042)	0.0591(0.0025)	0.0577(0.0031)
		GES	0.0430(0.0005)	0.9217(0.0024)	0.0865(0.0025)	0.0605(0.0025)
		PC	0.0241(0.0003)	0.6247(0.0149)	0.1301(0.0057)	0.2110(0.0092)
		NOTEARS	**	**	**	**
	500	NL	0.0295(0.0008)	0.6055(0.0142)	0.4839(0.0068)	0.4184(0.0084)
		NPVAR	**	**	**	**
		CAM	**	**	**	**
		RESIT	0.0368(0.0009)	0.8752(0.0056)	0.0994(0.0031)	0.0920(0.0040)
		GES	0.0394(0.0005)	0.8820(0.0035)	0.1190(0.0036)	0.0985(0.0036)
		PC	0.0247(0.0003)	0.6181(0.0141)	0.1774(0.0067)	0.2492(0.0097)
		NOTEARS	**	**	**	**

under the small n and p case in terms of HD and FDR. Notably, while the metric TPR of NL decreases as n increases in some scenarios of Examples 2 and 3, the other three metrics indicate that NL's performance improves in that HD and FDR decrease and MCC increases as n increases. This observation is largely due to the fact that when n is relatively small, many false directed edges are discovered resulting in a high value of FDR, and thus leads to a high value of TPR. As n increases, the estimation procedure in NL becomes more accurate and many false directed edges are eliminated leading to the decreasing of TPR. Note that NPVAR's performance o is less satisfactory, possible due to the fact that it is designed for the equal variance case and the considered additive modeling can not detect the interaction relations. It is also worth pointing out that NL's performance may be further improved with a finer tuning scheme at the cost of increasing computational cost. Note that NPVAR, CAM and PC do not produce any results for the cases with $p = 100$ in Examples 1 to 2 after more than 24 hours, which indicates that they may suffer serious computational burden even when dealing with a medium-sized DAG.

Moreover, we report the averaged running times of all the competing methods under various examples and scenarios in Table 4. Precisely, we consider the same generating schemes in Examples 1 to 3 and vary (n, p) from $(200, 50), (200, 100)$, $(500, 50)$ to $(500, 100)$. It is thus clear from Table 4 that, compared to all the other competitors, NL is remarkably computational efficient, especially when the number of nodes is relatively large. It is also interesting to notice that the averaged running times of GES are less than those of NL at the cost of achieving less satisfactory numerical results as illustrated in Tables 1 to 3.

5.2. Application to cell signalling data

In this section, we apply NL and all the other competitors to analyze the multivariate flow cytometry data from Sachs et al. (2005), which consists of continuous measurement of multiply phosphorylated proteins and phospholipid components following perturbation of thousands of individual human immune system cells with molecular interventions. Precisely, the intracellular signaling networks of human primary naive CD4+ T-cells are studied by recording cell reactions terminated by 15 fixation minutes after a series of interventions, and flow cytometry measurements are taken from 11 expression levels of proteins and phospholipids under 9 experimental conditions. Note that this data can be regarded as a common benchmark in causal inference, because it comes with a known consensus network and is widely accepted by the biological community.

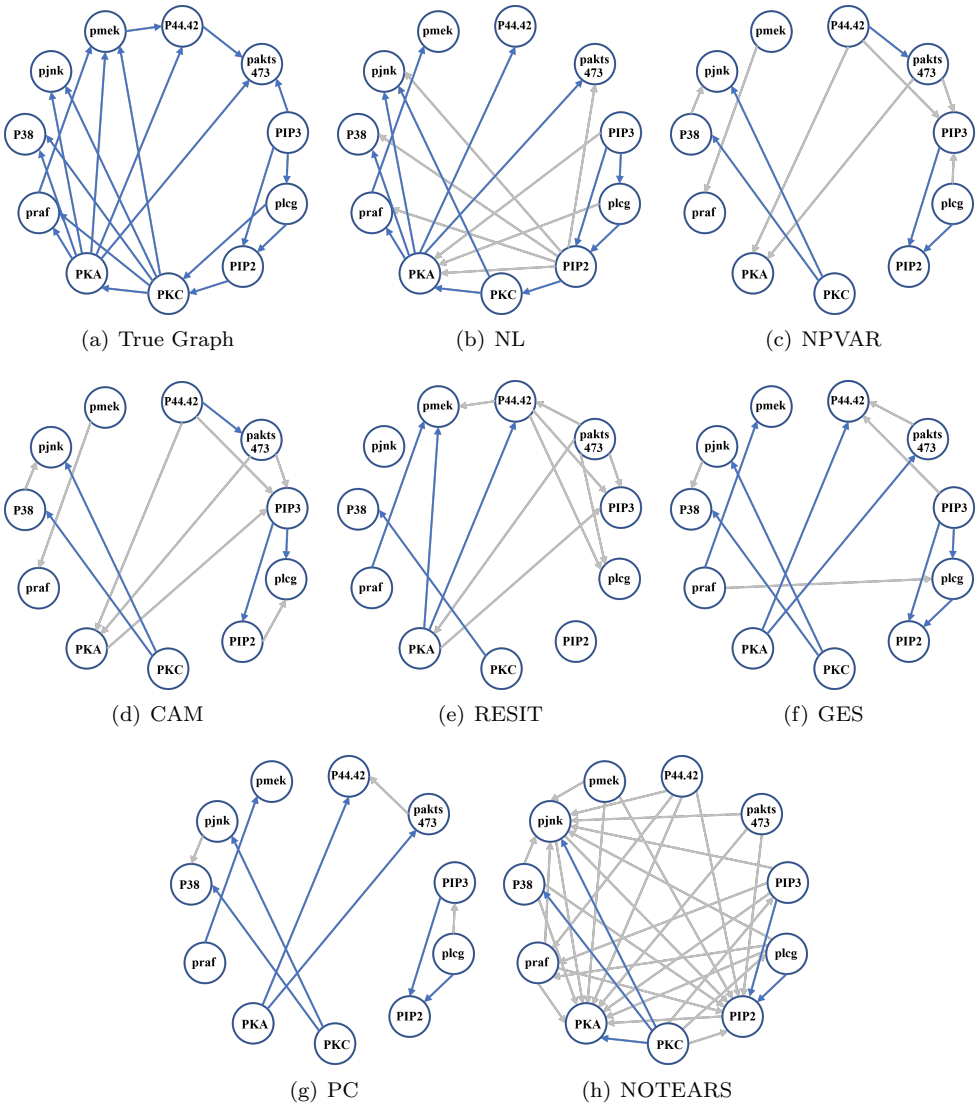
Following the same treatment as in Yuan et al. (2019), we also consider one specific condition among the 9 experimental conditions, which uses anti-cluster of differentiation 3 (CD3) /cluster of differentiation 28 (CD28) and intercellular adhesion molecule-2 (ICAM-2) as general perturbations since they attempt to activate cell signaling. Then, we use the consensus network in Sachs et al. (2005)

Table 4. The averaged running times (in minutes) of all the competitors under various scenarios of Examples 1 to 3 together with their standard errors in parentheses. Here, ** denotes that the corresponding methods did not produce any results after running for 24 hours.

n	p	Methods	Example 1	Example 2	Example 3
200	50	NL	4.30(0.13)	2.97(0.10)	3.92(0.48)
		NPVAR	**	**	**
		CAM	590.33(40.40)	590.19(47.06)	590.94(27.95)
		RESIT	25.37(0.21)	26.62(0.56)	25.22(0.53)
		GES	0.09(0.01)	0.07(0.01)	0.10(0.02)
		PC	3.73(1.15)	2.32(1.25)	0.08(0.01)
		NOTEARS	45.81(21.10)	21.95(8.37)	240.12(69.53)
	100	NL	7.70(0.21)	5.14(0.15)	7.76(0.71)
		NPVAR	**	**	**
		CAM	**	**	**
		RESIT	117.90(0.34)	120.95(0.52)	120.62(0.87)
		GES	0.61(0.13)	0.47(0.08)	0.98(0.14)
		PC	**	**	0.38(0.07)
		NOTEARS	393.86(120.93)	173.25(38.55)	2,441.87(641.96)
500	50	NL	29.32(0.37)	19.41(0.15)	27.45(2.47)
		NPVAR	38.89(3.12)	37.27(2.74)	109.26(8.70)
		CAM	**	**	**
		RESIT	186.52(1.96)	189.02(2.84)	181.36(4.61)
		GES	0.09(0.01)	0.07(0.01)	0.12(0.01)
		PC	1,225.07(576.60)	175.29(73.81)	0.09(0.01)
		NOTEARS	39.28(13.71)	18.62(4.56)	353.39(44.41)
	100	NL	55.45(0.16)	38.58(0.18)	56.08(6.28)
		NPVAR	**	**	**
		CAM	**	**	**
		RESIT	799.70(28.69)	811.64(11.94)	802.94(10.19)
		GES	0.73(0.21)	0.40(0.05)	0.81(0.13)
		PC	**	**	0.44(0.09)
		NOTEARS	176.17(43.89)	132.30(67.72)	2,679.03(668.18)

as the true network as illustrated in Figure 3(a), and apply all the competitors to analyze this data. The learned DAGs are illustrated in Figure 3 and the numerical metrics of all the competitors are also reported in Table 5.

Figure 3 clearly shows that NL is the best performer in that 19 directed edges are learned and among them, 12 agree with the true DAG and 7 are falsely reconstructed. GES is the second-best performer that 8 directed edges are correctly recovered and the next is PC with 7 edges correctly estimated. NPVAR and CAM perform similarly, they both correctly identify 5 edges and obtain similar skeleton structures, largely due to the fact that they both use



(a) True Graph (b) NL (c) NPVAR

(d) CAM (e) RESIT (f) GES

(g) PC (h) NOTEARS

Figure 3. The true DAG and DAG learned by the seven competitors. Correct discoveries are represented by solid blue/dark lines, false discoveries are displayed as solid grey/light lines.

additive modeling and thus may miss some directed relations with a more general dependence structure. NOTEARS also correctly recover 5 directed edges at the cost of returning a large number of false discovered edges. RESIT only correctly identifies 4 edges. Notably, many true edges, such as PKC $\rightarrow$ PKA and PKA $\rightarrow$ P38, are correctly recovered by NL, but are missed by all the other methods. This indicates that NL can detect more general causal relations among collected nodes. Its superior performance is also supported by the numerical metrics in Table 5, where it excels under almost all the evaluation metrics.

Table 5. The numerical performance of all the competitors in application to cell signalling data.

Methods	SHD	FDR	TPR	MCC
NL	0.1364	0.3684	0.60	0.5418
NPVAR	0.2000	0.5833	0.25	0.2246
CAM	0.2091	0.6154	0.25	0.2049
RESIT	0.2182	0.6667	0.20	0.1501
GES	0.1455	0.3333	0.40	0.4479
PC	0.1455	0.3000	0.35	0.4321
NOTEARS	0.3909	0.8485	0.25	-0.0227

6. Discussion

This paper proposes an efficient method to learn nonparametric DAG from observed data with sound statistical guarantees. It leverages the concept of topological layers to facilitate nonparametric DAG learning , connecting it with kernel ridge regression and learning gradients by showing that the introduced layers can be exactly reconstructed via kernel ridge regression. More interestingly, the parent-child relations can be simultaneously recovered without any extra estimation by using the derivative reproducing property in the smooth RKHS. An efficient learning algorithm is developed and the statistical guarantees of the proposed method in terms of exact DAG recovery are established ensuring the underlying DAG with general parent-child dependence can be exactly recovered. Its superior performance is also supported by numerical experiments on various simulated and real-life examples. It is worth noting that one of the possible future work is to modify the proposed method to reconstruct the topological layers in a bottom-up fashion if some decreasing noise-variance assumption is satisfied, where the topological layers should be defined based on the longest distance to one of the leaf nodes.

Supplementary Material

The supplementary material includes all the theoretical proofs.

Acknowledgments

The authors thank the associate editor and three anonymous referees for their constructive suggestions, which significantly improved this paper. He's research was supported in part by NSFC-11901375, Shanghai Research Center for Data Science and Decision Technology, Fundamental Research Funds for the Central Universities, and Program for Innovative Research Team of Shanghai University of Finance and Economics. Lv's research was partially supported by NSFC-12371291 and 72342019, Qinglan project of JiangSu Province (2022).

References

Bühlmann, P., Peters, J. and Ernest, J. (2013). CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* **42**, 2526–2556.

Chen, W., Drton, M. and Wang, Y. (2019). On causal discovery with an equal-variance assumption. *Biometrika* **106**, 973–980.

Chickering, D. M. (2003). Optimal structure identification with greedy search. *Journal of Machine Learning Research* **3**, 507–554.

Durrande, N., Ginsbourger, D., Roustant, O. and Carraro, L. (2013). ANOVA kernels and RKHS of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis* **115**, 57–67.

Gao, M., Ding, Y. and Aragam, B. (2020). A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems* **33**, 11599–11611.

Ghoshal, A. and Honorio, J. (2018). Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics* **PMLR 84**, 1466–1475.

He, T., Zhong, P., Cui, Y. and Mandrekar, V. (2022). Unified tests for nonparametric functions in RKHS with kernel selection and regularization. *Statistica Sinica* **33**, 919–944.

He, X., Wang, J. and Lv, S. (2021). Efficient kernel-based variable selection with sparsistency. *Statistica Sinica* **31**, 2123–2151.

Hyvärinen, A. and Smith, S. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research* **14**, 111–152.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8**, 613–636.

Li, C., Shen, X. and Pan, W. (2020). Likelihood ratio tests for a large directed acyclic graph. *Journal of the American Statistical Association* **115**, 1304–1319.

Li, C., Shen, X. and Pan, W. (2023). Nonlinear causal discovery with confounders. *Journal of the American Statistical Association* **119**, 1205–1214.

Lindsay, B., Markatou, M., Ray, S., Yang, K. and Chen, S. (2008). Quadratic distances on probabilities: A unified foundation. *The Annals of Statistics* **36**, 983–1006.

Lv, S., Lin, H., Lian, H. and Huang, J. (2018). Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space. *The Annal of Statistics* **46**, 781–813.

Mooij, J., Peters, J., Janzing, D., Zscheischler, J. and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research* **17**, 1103–1204.

Park, G. (2020). Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research* **21**, 2896–2929.

Park, G. and Raskutti, G. (2018). Learning quadratic variance function (QVF) DAG models via overdispersion scoring (ODS). *Journal of Machine Learning Research* **18**, 8300–8342.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101**, 219–228.

Peters, J., Janzing, D. and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA.

Peters, J., M. Mooij, J., Janzing, D. and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research* **15**, 2009–2053.

Rosasco, L., Villa, S., Mosci, S., Santoro, M. and Verri, A. (2013). Nonparametric sparsity and regularization. *Journal of Machine Learning Research* **14**, 1665–1714.

Rothenhäusler, D., Ernest, J. and Bühlmann, P. (2018). Causal inference in partially linear

structural equation models. *The Annals of Statistics* **46**, 2904–2938.

Sachs, K., D. Perez, O., Pe'er, D., A. Lauffenburger, D. and P. Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529.

Shimizu, S., Hyvärinen, A. and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7**, 2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T. et al. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research* **12**, 1225–1248.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018). Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *Journal of Machine Learning Research* **19**, 1–29.

Smale, S. and Zhou, D. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation* **26**, 153–172.

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search.* MIT Press, Cambridge, MA.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machine*. Springer.

Steinwart, I. and Scovel, C. (2011). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation* **35**, 363–417.

Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13**, 689–705.

Sun, W., Wang, J. and Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research* **14**, 3419–3440.

Tsamardinos, I., Brown, L. and Aliferis, C. (2006). The Max-Min Fill-Climbing Bayesian network structure learning algorithm. *Machine Learning* **65**, 31–78.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In *Advances in Kernel Methods: Support Vector Learning*, 69–88. MIT Press, Cambridge, MA.

Wang, Y. and Drton, M. (2020). High-dimensional causal discovery under non-Gaussianity. *Biometrika* **107**, 41–59.

Yuan, Y., Shen, X., Pan, W. and Wang, Z. (2019). Constrained likelihood for reconstructing a directed acyclic Gaussian graph. *Biometrika* **106**, 109–125.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence* **25**, 647–655.

Zhang, K., Wang, Z., Zhang, J. and Schölkopf, B. (2016). On estimation of functional causal models: General results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology* **7**, 1–22.

Zhang, X., Dan, C., Aragam, B., Ravikumar, P. and Xing, E. (2020). Learning sparse nonparametric DAGs. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* **PMLR 108**, 3414–3425.

Zhao, R., He, X. and Wang, J. (2022). Learning linear non-Gaussian directed acyclic graph with diverging number of nodes. *Journal of Machine Learning Research* **23**, 12314–12347.

Zhou, D. (2007). Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics* **220**, 456–463.

Zhou, W., He, X., Zhong, W. and Wang, J. (2022). Efficient learning of quadratic variance function directed acyclic graphs via topological layers. *Journal of Computational and Graphical Statistics* **31**, 1269–1279.

(Received August 2022; accepted June 2023)