# Optimal Kernel Choice for Score Function-based Causal Discovery

**Wenjie Wang** [1 2] **Biwei Huang** [3] **Feng Liu** [4] **Xinge You** [5] **Tongliang Liu** [6] **Kun Zhang** [2 7] **Mingming Gong** [1 2]

## Abstract

Score-based methods have demonstrated their effectiveness in discovering causal relationships by scoring different causal structures based on their goodness of fit to the data. Recently, Huang et al. (2018) proposed a generalized score function that can handle general data distributions and causal relationships by modeling the relations in reproducing kernel Hilbert space (RKHS). The selection of an appropriate kernel within this score function is crucial for accurately characterizing causal relationships and ensuring precise causal discovery. However, the current method involves manual heuristic selection of kernel parameters, making the process tedious and less likely to ensure optimality. In this paper, we propose a kernel selection method within the generalized score function that automatically selects the optimal kernel that best fits the data. Specifically, we model the generative process of the variables involved in each step of the causal graph search procedure as a mixture of independent noise variables. Based on this model, we derive an automatic kernel selection method by maximizing the marginal likelihood of the variables involved in each search step. We conduct experiments on both synthetic data and real-world benchmarks, and the results demonstrate that our proposed method outperforms heuristic kernel selection methods.

[1]School of Mathematics and Statistics, The University of Melbourne, Australia [2]Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates [3]Halicioğlu Data Science Institute (HDSI), University of California, San Diego, United States [4]School of Computing and Information Systems, The University of Melbourne, Australia [5]Huazhong University of Science and Technology, China [6]School of Computer Science, Faculty of Engineering, The University of Sydney, Australia [7]Department of Philosophy, Carnegie Mellon University, United States. Correspondence to: Mingming Gong <mingming.gong@unimelb.edu.au>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

## 1. Introduction

Understanding causal structures is a fundamental scientific problem that has been extensively explored in various disciplines, such as social science (Antonakis & Lalive, 2011), biology (Londei et al., 2006), and economics (Moneta et al., 2013). While conducting randomized experiments are widely regarded as the most effective method to identify causal structures, their application is often constrained by ethical, technical, or cost-related reasons (Spirtes & Zhang, 2016). Consequently, there is a pressing need to develop causal discovery methods that can infer causal structures only from uncontrolled observational data. In recent years, score-based methods have emerged as a promising approach, enabling significant advancements in the field of causal discovery (Zhang et al., 2018; Vowels et al., 2022).

Score-based methods are commonly employed for causal discovery by evaluating candidate causal graphs based on the specific criteria. Various search strategies, including traditional discrete search-based methods (Chickering, 2002; Silander & Myllymäki, 2006; Yuan & Malone, 2013) and the recent continuous optimization-based methods (Zheng et al., 2018; Yu et al., 2019; Zhu et al., 2019; Zheng et al., 2020), are then utilized to search for the graph with the optimal score. Therefore, score-based methods require assuming specific statistical models for causal relationships and data distributions, such as the BIC score (Schwarz, 1978), MDL score (Adriaans & van Benthem, 2008) and BGe score (Geiger & Heckerman, 1994) only for linear-Gaussian models and some for other explicit model classes (Bühlmann et al., 2014; Hyvärinen & Smith, 2013; Sokolova et al., 2014). And such assumptions may limit their applicability in real-world scenarios.

Recently, Huang et al. (2018) proposed a generalized score function that can handle diverse causal relationships and data distributions. This approach utilizes conditional cross-covariance operator (Fukumizu et al., 2004) to transform the general conditional independence test into a regression model selection problem. The proposed score function assumes a causal relationship in RKHS, where the response variable are first projected through a pre-defined kernel function. And then, it is assumed that these fixed features are generated following a nonlinear additive model from their parent variables. Although their proposed score function

can detect general relationships, their results are significantly influenced by the selected kernel parameters. Like most kernel-based methods, Huang et al. (2018) opted for a heuristic kernel parameter strategy, where the bandwidth of the kernel is chosen to be the median of the inter-sample distance. However, such median heuristic does not take into account the inherent characteristics of the involved variables, potentially failing to capture the correct relationships.

Kernel selection is a fundamental problem in kernel-involved tasks (Chu & Marron, 1991; Herrmann et al., 1992; Chapelle & Vapnik, 1999; Kim et al., 2006). In practice, this choice often reduces to the calibration of the bandwidth, which may even be more important than the choice of the kernel families (Schölkopf et al., 2002, Section 4.4.5). Even though the median heuristic has been extensively employed in various applications (Gretton et al., 2005; 2012a; Zhang et al., 2011; Liu et al., 2016), it has no guarantees of optimality (Ramdas et al., 2015; Garreau et al., 2017). Several methods for kernel bandwidth selection have been proposed, including selecting the kernel by maximizing the test power in the two-sample test problem (Gretton et al., 2012b; Liu et al., 2020; Sutherland et al., 2021). However, the optimal kernel choice, influenced by the specific task and inherent data nature, lacks a unified criterion, making existing kernel selection methods unsuitable to the generalized score mentioned above for causal discovery.

**Contributions.** In this paper, we propose a kernel selection method within the kernel-based generalized score functions for causal discovery. Instead of using the median heuristic, our method can automatically select the optimal kernel bandwidth that best fit the given data. To achieve this, we extend the RKHS regression framework and model the generative process of the involved variables in the graph as a mixture of independent noise variables. Correspondingly, we estimate the causal relationships of the involved variables by maximizing the marginal likelihood of their joint distributions. In our approach, the parameters of kernels is automatically learned from data, along with other parameters in the model. Consequently, our method can effectively select the optimal kernel parameters, which is superior to the median heuristic-based ones. We conduct experiments on both synthetic datasets and real-world benchmarks. The experimental results demonstrate the effectiveness of our approach in selecting the kernel bandwidth while improving the accuracy of causal relationship discovery.

## 2. Background

The general conditional independence can be characterized in RKHS by conditional cross-covariance operator without the assumptions on causal mechanisms and data distributions. We first give some notations and introduce some properties that will be used throughout of the paper. We take $(\Omega, \mathcal{F}, P)$ as the underlying probability space. Let $(\mathcal{X}, \mathscr{X})$, $(\mathcal{Y}, \mathscr{Y})$ and $(\mathcal{Z}, \mathscr{Z})$ be separable measurable spaces, and let $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ and $Z : \Omega \to \mathcal{Z}$ be random variables with distributions $P_X$, $P_Y$ and $P_Z$.

### 2.1. Conditional Cross-covariace Operator on RKHS

For the random variable $X$ on $\mathcal{X}$, we define a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\mathcal{X}$ on $\mathcal{X}$ with a reproducing kernel $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which satisfies 1. $\forall x \in \mathcal{X}, k_\mathcal{X}(x, \cdot) \in \mathcal{H}_\mathcal{X}$; 2. $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{H}_\mathcal{X}, f(x) = \langle f, k_\mathcal{X}(x, \cdot)\rangle_{\mathcal{H}_\mathcal{X}}$. And the corresponding canonical feature map $\phi_\mathcal{X} : \mathcal{X} \to \mathcal{H}_\mathcal{X}$ has the form $\phi_\mathcal{X}(x) = k_\mathcal{X}(x, \cdot)$ with $\langle \phi_\mathcal{X}(x), \phi_\mathcal{X}(x')\rangle_{\mathcal{H}_\mathcal{X}} = k_\mathcal{X}(x, x')$ (Steinwart & Christmann, 2008, Lemma 4.19). Similarly we define $k_\mathcal{Y}, \phi_\mathcal{Y}(y), \mathcal{H}_\mathcal{Y}, k_\mathcal{Z}, \phi_\mathcal{Z}(z), \mathcal{H}_\mathcal{Z}$.

Following the definition of cross-covariance operator $\Sigma_{XY}$ (Baker, 1973), Fukumizu et al. (2004) define the conditional cross-covariance operator $\Sigma_{XX|Y}$. For the random vector $(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$, $\Sigma_{XX|Y}$ captures the conditional variance in the following way:

$$\left\langle g, \Sigma_{XX|Y} g\right\rangle_{\mathcal{H}_\mathcal{X}} = E_Y[\mathrm{Var}_{X|Y}[g(X)|Y]]. \tag{1}$$

where $g$ is an arbitrary function in $\mathcal{H}_\mathcal{X}$. The conditional cross-covariance operator is related to the conditional independence according to the following lemma.

**Lemma 2.1.** *(Fukumizu et al., 2004, Theorem 7) Let $(\mathcal{H}_\mathcal{X}, k_\mathcal{X}), (\mathcal{H}_\mathcal{Y}, k_\mathcal{Y})$ and $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ be reproducing kernel Hilbert spaces over measurable spaces $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$, with continuous and bounded kernels. Let $X$ , $Y$ and $Z$ be random variables on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$. Denote $\ddot{Y} = (Y, Z)$ and $\mathcal{H}_{\ddot{\mathcal{Y}}} = \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Z}$ be the direct product. It is assumed that $E_{X|Y}[g(X) \mid Y = \cdot] \in \mathcal{H}_\mathcal{Y}$ and $E_{X|\ddot{Y}}\left[g(X) \mid \ddot{Y} = \cdot\right] \in \mathcal{H}_{\ddot{\mathcal{Y}}}$ for all $g \in \mathcal{H}_\mathcal{X}$. Then*

$$\Sigma_{XX|Y} \geq \Sigma_{XX|\ddot{Y}}. \tag{2}$$

*And if further $k_\mathcal{X}$ is **characteristic** kernel (Fukumizu et al., 2007), the following equation holds:*

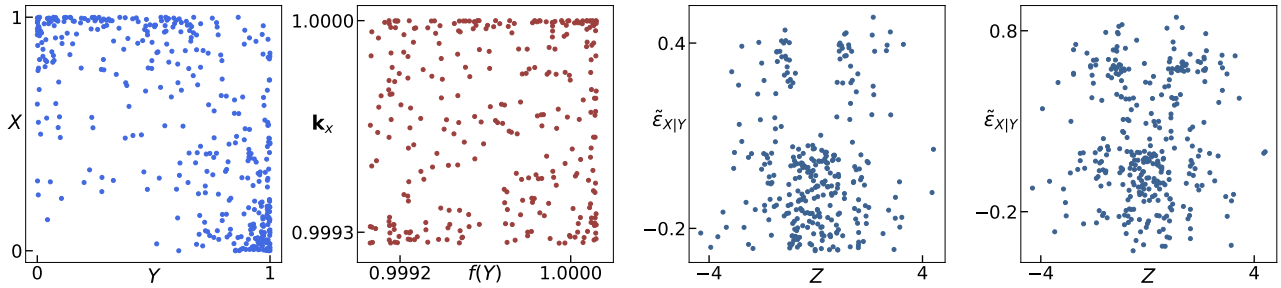
$$\Sigma_{XX|Y} - \Sigma_{XX|\ddot{Y}} = 0 \Longleftrightarrow X \perp\!\!\!\perp Y \mid Z. \tag{3}$$

From these relations, one can identify the parent variables of a given variable from the sets of variables, which has been explored in (Fukumizu et al., 2009).

### 2.2. Regression in RKHS

Based on the above lemma, the conditional independence test can be transformed into a regression model selection problem. By combining Eq. 1 and Eq. 3, we have

$$\begin{aligned} E_Y[\mathrm{Var}_{X|Y}[g(X) \mid Y]] &= E_{\ddot{Y}}[\mathrm{Var}_{X|\ddot{Y}}[g(X) \mid \ddot{Y}]] \\ &\Longleftrightarrow X \perp\!\!\!\perp Y \mid Z, \text{for all } g \in \mathcal{H}_\mathcal{X}. \end{aligned} \tag{4}$$




(a) Original data (b) Feature space (c) HSIC = 0.0062 (Median) (d) HSIC = 0.0038 (Ours)

*Figure 1.* Visualization of features and estimated noise using median heuristic or learnable bandwidth. (a) Scatter plot of original data $Y$ and $X$. (b) Scatter plot of the projected features $f(Y)$ and $\boldsymbol{k}_x$ using conditional likelihood-based score with **trainable** $k_{\mathcal{X}}$. (c) Scatter plot of the estimated regression noise $\tilde{\varepsilon}_{X|Y}$ and $Z$ with median heuristic bandwidth. We utilized HSIC to quantify the independence between $\tilde{\varepsilon}_{X|Y}$ and $Z$. A lower HSIC value indicates a higher degree of independence between them. (d) Scatter plot of $\tilde{\varepsilon}_{X|Y}$ and $Z$ with trainable $k_{\mathcal{X}}$ using our proposed score function (Ours).

Let us consider the following regression functions in RKHS:

$$\begin{aligned} \phi_{\mathcal{X}}(X) &= f_1(Y) + \varepsilon_1, \\ \phi_{\mathcal{X}}(X) &= f_2(\ddot{Y}) + \varepsilon_2, \end{aligned} \tag{5}$$

where $\phi_{\mathcal{X}}(X)$ is the continuous feature mapping on $\mathcal{H}_{\mathcal{X}}$, $\varepsilon_1$ and $\varepsilon_2$ are the independent noise. Since each orthogonal component in $\phi_{\mathcal{X}}(X)$ can be replaced with a $g \in \mathcal{H}_{\mathcal{X}}$, then

$$\begin{aligned} E_Y[\mathrm{Var}_{X|Y}[\phi(X) \mid Y]] &= E_{\ddot{Y}}[\mathrm{Var}_{X|\ddot{Y}}[\phi(X) \mid \ddot{Y}]] \\ &\Longleftrightarrow X \perp\!\!\!\perp Y \mid Z. \end{aligned} \tag{6}$$

We can utilize Eq. 6 to transform the conditional independence test into a model selection problem. That is, within the regression model in RKHS (Eq. 5), given a considering random variable $X$, the model with the best fit will be chosen. Furthermore, if the models have equivalent fitting performance, we will opt for the model with the minimum population of independent variables.

## 3. Motivation

In line with the aforementioned theory, Huang et al. (2018) introduced a score-based causal discovery method capable of investigating general conditional independence relations without prior information on relational models and data. Generally, it employs conditional likelihood to assess the goodness of fit of data within the regression model in the RKHS. Specifically, for a random variable $X$ and its independent variables $PA$ on the domains $\mathcal{X}$ and $\mathcal{P}$ respectively, we define the corresponding RKHS $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ on $\mathcal{X}$, utilizing the feature map $\phi_{\mathcal{X}}$ as per previous notation. It assumes a causal relation in the RKHS, where the feature $\phi_{\mathcal{X}}(X)$ is generated as follows,

$$\phi_{\mathcal{X}}(X) = f(PA) + \boldsymbol{\varepsilon}, \tag{7}$$

where $f : \mathcal{P} \to \mathcal{H}_{\mathcal{X}}$ represents a nonlinear mapping and $\boldsymbol{\varepsilon} \in \mathcal{H}_{\mathcal{X}}$ denotes independent noise.

Since $\phi_{\mathcal{X}}(X)$ is in the infinite-dimensional $\mathcal{H}_{\mathcal{X}}$, which is hard to measure, with the property that functions in the RKHS are in the closure of linear combinations of the kernel at given points, in practice, we can project $\phi_X(X)$ onto a empirical vector form without losing information (Schölkopf et al., 2002, Section 2.2.6). Suppose we have $n$ observations $(\boldsymbol{x}, \boldsymbol{pa}) = [(x^1, pa^1), (x^2, pa^2), \cdots, (x^n, pa^n)]$, the empirical form of $\phi(x)$ is represented by $\boldsymbol{k}_x = \left[k_{\mathcal{X}}(x, x^1), \cdots, k_{\mathcal{X}}(x, x^n)\right]^\top$ with the empirical feature space, denoted as $\mathcal{F}_{\mathcal{X}}$. Then, the regression in RKHS on finite observations is reformulated as

$$\boldsymbol{k}_x = f(pa) + \boldsymbol{\varepsilon}, \tag{8}$$

where $f : \mathcal{P} \to \mathcal{F}_{\mathcal{X}}$, $\boldsymbol{\varepsilon} = [\varepsilon_1, \cdots, \varepsilon_n]^T \in \mathcal{F}_{\mathcal{X}}$ and $\boldsymbol{k}_x$ is determined by the parameters in $k_{\mathcal{X}}$. Therefore, the choice of parameters in $k_{\mathcal{X}}$ significantly influences the model's fitting performance. However, in Huang et al. (2018)'s method, a common heuristic kernel selection strategy is employed. Specifically, it opts for the Gaussian kernel and sets the bandwidth as twice the median distance between data points. In other words, in Huang et al. (2018)'s method, the feature of response variable $\boldsymbol{k}_x$ is pre-defined and fixed, correspondingly using conditional likelihood as the score function. While such kernel bandwidth selection is straightforward, it has no guarantees of optimality. Due to the lack of consideration for the inherent characteristics of the data, such median heuristic may result in inaccurate assessments of the independence relationships among variables. we will start with a simple example demonstrating the spurious result caused by the heuristic kernel width selection method, as depicted in Figure 1.

Let us consider a scenario where the random variables $Z, Y$ and $X$ satisfy: $Z = E_1$, $Y = \cos(1.5Z^2 + E_2)^2$ and $X = \cos(1.5Y^2 + E_3)^2$, where $E_1 \sim \mathcal{N}(0, 1)$ and $E_2, E_3 \sim \mathcal{N}(0, 0.5)$. Figure 1(a) displays the original data of $Y$ and $X$, which are dependent. However, adhering to

the heuristic kernel selection strategy and fixing the parameters of the kernel function $k_{\mathcal{X}}$ can unexpectedly lead to a spurious edge $Z \to X$, incorrectly suggesting that $Z$ is a parent node of $X$. This is attributed to the inappropriate bandwidth choice of $k_{\mathcal{X}}$ is median heuristically selected. Consequently, the model fails to adequately capture the relationship between $Y$ and the fixed response features $\boldsymbol{k}_x$. Figure 1(c) depicts the scatter plot of the estimated regression noise $\tilde{\varepsilon}_{X|Y} = \boldsymbol{k}_x - f(y)$ and $Z$. With $n$ observations, the estimated empirical noise $\tilde{\varepsilon}_{X|Y}$ is a $n$-dimensional vector, and Figure 1(c) only illustrates the relationship between the first dimension of $\tilde{\varepsilon}_{X|Y}$ and $Z$. From Figure 1(c), it is evident that there remains a correlation between the estimated noise $\tilde{\varepsilon}_{X|Y}$ and $Z$, which prompts the model to mistakenly consider $Z$ as a parent node of $X$.

To select appropriate kernel parameters based on the inherent nature of the given data, an alternative way is to make the parameters of $k_{\mathcal{X}}$ trainable, allowing them to be learned alongside other parameters in the regression model. However, the existing score function in (Huang et al., 2018) is based on the likelihood of conditional distribution $p(\boldsymbol{k}_x \mid y; f)$, inherently lacking constraints on the parameters of $k_{\mathcal{X}}$. If we continue to use such a conditional likelihood-based score function with the trainable kernel parameters, the model will end up learning a trivial constant mapping with an excessively high score, which is uninformative for causal discovery purposes. Figure 1(b) illustrates the first dimension of the learned empirical features $\boldsymbol{k}_x$ and the projected feature $f(y)$ using the conditional marginal likelihood with trainable parameters. Notably, all the features are projected into an extremely small space.

By illustrating the potential limitations of the heuristically-selected kernels, we emphasize the necessity for an automated kernel parameter selection method tailored to the specific data and a novel score function to improve the accuracy of estimating causal relationships. In the following section, we will introduce our kernel selection method, which can automatically select the optimal kernel parameters for the given data through optimization. Figure 1(d) illustrates the scatter plot of $Z$ and the estimated noise learned through our method, highlighting the increased independence between them. For a more precise comparison, we employed HSIC (Gretton et al., 2005) to quantify the independence between $\tilde{\varepsilon}_{X|Y}$ and $Z$. And a lower HSIC value indicates greater independence between them. Compared with the median heuristic-based approach (Figure 1.c), the lower HSIC observed in our method suggests that our proposed score function with the trainable kernel parameters can better fit the given data, thereby correctly blocking the influence of $Z$ on $X$. For more experiment details, please refer to Appendix B.3.3.

# 4. Optimal Kernel Selection via Minimizing Mutual Information

In this section, we present a mutual information-based kernel selection method which can automatically select the optimal kernel for the given data. We first extend the regression model in RKHS by replacing the original pre-fixed kernel with a trainable one and consider the causal relationship as a mixture of independent noise variables. Hence, we minimize the mutual information between the estimated noise and the independent variables to simultaneously learn the parameters of both the kernel function $k_{\mathcal{X}}$ and the projection $f$. In a learning manner, our method can automatically select the optimal kernel parameters that best fits the given data while maintaining simplicity and straightforwardness.

## 4.1. Preliminaries

Suppose $X$ and $PA$ are one of the random variables and its parents in the given DAG $\mathcal{G}$, with the domain $\mathcal{X}$ and $\mathcal{P}$ respectively. We denote the RKHS $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ on $\mathcal{X}$ with the continuous feature mapping $\phi_{\mathcal{X}}(X)$ as before. Similarly we define $(\mathcal{H}_{\mathcal{P}}, k_{\mathcal{P}})$ on $\mathcal{P}$. Suppose we have $n$ observations $(\boldsymbol{x}, \boldsymbol{pa}) = [(x^1, pa^1), \cdots, (x^n, pa^n)]$, where both $x^i$ and $pa^i$ may have more than one dimension. We define $K_X, K_{PA} \in \mathbb{R}^{n\times n}$ are kernel matrices containing entries $K_{X(ij)} = k_{\mathcal{X}}(x^i, x^j)$ and $K_{PA(ij)} = k_{\mathcal{P}}(pa^i, pa^j)$.

## 4.2. Mutual information-based score function

To effectively constrain the parameters in $k_{\mathcal{X}}$, we propose using the mutual information between the given $X$ and its potential parent nodes $PA$ as the score function, which can simultaneously supervise all the trainable parameters in our extended model. For a specific observation $(x, pa)$, the empirical feature vector of $x$ is represented as $\boldsymbol{k}_x = \left[k_{\mathcal{X}}(x^1, x), k_{\mathcal{X}}(x^2, x), \cdots, k_{\mathcal{X}}(x^n, x)\right]^T$. According to Eq. 8, if the assumed relation holds, there exist $k_{\mathcal{X}}$ and $f$ such that the noise $\boldsymbol{\varepsilon}$ is independent of $PA$. Therefore, the parameters of $k_{\mathcal{X}}$ and $f$ can be learned by ensuring that $PA$ and the estimated $\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{k}_x - f(pa)$ are as independent as possible. This is achieved by minimizing their mutual information:

$$I(PA, \tilde{\boldsymbol{\varepsilon}}) = H(PA) + H(\tilde{\boldsymbol{\varepsilon}}) - H(PA, \tilde{\boldsymbol{\varepsilon}}), \tag{9}$$

where the joint entropy $H(PA, \tilde{\boldsymbol{\varepsilon}}) = -\log p(PA, \tilde{\boldsymbol{\varepsilon}})$ and $\tilde{\boldsymbol{\varepsilon}} = [\tilde{\varepsilon}_1, \cdots, \tilde{\varepsilon}_n]^\top \in \mathcal{F}_{\mathcal{X}}$. With each $\tilde{\varepsilon}_i$ in $\tilde{\boldsymbol{\varepsilon}}$ being independent of the other dimensions, the joint distribution can be decomposed as $p(PA, \tilde{\boldsymbol{\varepsilon}}) = \prod_i p(PA, \tilde{\varepsilon}_i)$. For each dimension, we employ the change-of-variables theorem to obtain the following form

$$p(PA, \tilde{\varepsilon}_i) = p(PA, X)/|\det \mathbf{J}_i|, \tag{10}$$

where $\mathbf{J}_i$ is the Jacobian matrix of transformation from $(PA, X)$ to $(PA, \tilde{\varepsilon}_i)$, i.e. $\mathbf{J}_i = [\partial(PA, \tilde{\varepsilon}_i)/\partial(PA, X)]$.

Therefore, $H(PA, \tilde{\varepsilon})$ is derived as

$$\begin{aligned} H(PA, \tilde{\varepsilon}) &= -E\left[\sum_i \log p(PA, \tilde{\varepsilon}_i)\right] \\ &= nH(PA, X) + \sum_i E[\log|\det \mathbf{J}_i|], \end{aligned} \tag{11}$$

By combining Eq.9 and Eq.11, we can derive that

$$I(PA, \tilde{\varepsilon}) = -\sum_i (E[\log p(\tilde{\varepsilon}_i)] + E[\log|\det \mathbf{J}_i|]) + C, \tag{12}$$

where $C = H(PA) - nH(PA, X)$ does not depend on the parameters in $k_{\mathcal{X}}$ and $f$ and can be considered as constant. Hence, we focus only on the first two term in Eq. 12, which are equal to the negative log-likelihood of the joint distribution $p(X, PA)$ when the independent noise $\varepsilon$ is assumed to follow an isotropic Gaussian distribution. (Refer to Appendix A.1 for the detailed derivations.)

Suppose $(x^j, pa^j) \in (\mathcal{X}, \mathcal{P})$ is one of the $n$ observations in $(\boldsymbol{x}, \mathbf{pa})$, where $PA$ has $m$ dimensions and $X$ has $l$ dimensions ($l > 1$ when $X$ is a multi-dimensional variable). Minimizing the mutual information $I(PA, \tilde{\varepsilon})$ in our model is equivalent to minimizing the negative log likelihood of the joint distribution $p(PA, X)$, which is

$$\begin{aligned} l(f, \sigma_x) &= -\frac{1}{n}\sum_{j=1}^{n} \log p(x^j, pa^j \mid f, \sigma_x) \\ &= -\frac{1}{n}\sum_{i,j}\left[\log p(\tilde{\varepsilon}_i^j) + \log|\det \mathbf{J}_i^j|\right], \end{aligned} \tag{13}$$

where $\sigma_x$ is the parameters in $k_{\mathcal{X}}$ and $\tilde{\varepsilon}_i^j$ represents the $i$-th dimension of the estimated noise from observation $(x^j, pa^j)$, $\mathbf{J}_i^j$ is its Jacobian matrix, which can be calculated as

$$\begin{aligned} \mathbf{J}_i^j &= \begin{bmatrix} \frac{\partial PA}{\partial PA} & \frac{\partial PA}{\partial X} \\ \frac{\partial \tilde{\varepsilon}_i}{\partial PA} & \frac{\partial \tilde{\varepsilon}_i}{\partial X} \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{I}_m & \mathbf{0}_{m\times l} \\ -f'^{\top}(pa^j) & K'^{\top}_{X(j,i)} \end{bmatrix}_{(m+1)\times(m+l)}, \end{aligned} \tag{14}$$

where $m$ and $l$ are the dimensions of $PA$ and $X$ respectively, $\mathbb{I}_m$ is an $m$-th order identity matrix, $f'(pa^j)$ is the $n$-dimensional vector of the derivatives of $f$ at $pa^j$. And $K'_{X(j,i)}$ is $l$-dimensional derivative vector of $K'_{X(j,i)}$ with respect to $x^j = [x_1^j, \cdots, x_l^j]$, with the $t$-th element given by

$$K'_{X(j,i)}(t) = \frac{\partial k_{\mathcal{X}}(x^j, x^i)}{\partial x_t^j}. \tag{15}$$

It is obvious that $\mathbf{J}_i^j$ is a full row rank matrix, and its determinant can be replaced by its volume (Ben-Israel, 1999), which is expressed as

$$|\det \mathbf{J}_i^j| = \sqrt{\det \mathbf{J}_i^j \mathbf{J}_i^{j\top}} = \sqrt{\sum_t K'^2_{X(j,i)}(t)}. \tag{16}$$

See Appendix A.2 for the detailed derivations. From Eq. 16, it can be observed that the determinant is independent of $f$ but depends on the parameters of $k_{\mathcal{X}}$. In particular, when $X$ is a one-dimensional variable, $\mathbf{J}_i^j$ degenerates into an $(m+1)$-order lower triangular matrix, which is consistent with the determinant of square matrix.

On the other hand, it has been prone that employing models with a large number of parameters, such as multilayer perceptrons (MLPs), to approximate $f$ may lead to potential overfitting when dealing with a limited number of samples (compared to the complexity of the function class of $f$), especially in the post-nonlinear model (Uemura et al., 2022; Keropyan et al., 2023). With a trainable kernel $k_{\mathcal{X}}$, our model also faces similar challenges. To mitigate it, we opt to use a Gaussian process (Williams & Rasmussen, 1995) to model $f$ and accordingly use the marginal likelihood for supervision. That is, we use the marginal likelihood of the joint distribution $p(PA, X)$ as our score function. Specifically, we assume that $f \sim \mathcal{GP}(0, K_{PA})$, where $K_{PA}$ is the covariance matrix of $PA$. We then parameterize $f$ by integrating the parameters in $f$. For the random variable $X$ and its parent nodes $PA$ with the observation $(\boldsymbol{x}, \boldsymbol{pa})$, our score function, utilizing the marginal likelihood of joint density, is as follows:

$$\begin{aligned} S(X, PA) = &-\frac{1}{2}\mathrm{trace}\left\{K_X(K_{PA} + \sigma_\varepsilon^2 I)^{-1}K_X\right\} \\ &-\frac{n}{2}\log|K_{PA} + \sigma_\varepsilon^2 I| - \frac{n^2}{2}\log 2\pi \\ &+\sum_{i\neq j}^{n}\log\sqrt{\sum_t K'^2_{X(j,i)}(t)}, \end{aligned} \tag{17}$$

where $K_X$ and $K_{PA}$ are the kernel matrices and $\sigma_\varepsilon$ is the regularization parameter. See Appendix A.3 for the detailed derivations. The score function closely resembles the conditional marginal likelihood proposed in (Huang et al., 2018), differing notably in the inclusion of the last determinant term. However, it is precisely the last term that effectively supervises the kernel parameters, making the model to effectively fit the true relationships, rather than learning an uninformative consistent mapping.

Overall, for the current hypothetical DAG $\mathcal{G}$ with $q$ variables and observations $D = (\boldsymbol{x}, \boldsymbol{pa})$, our score function using the marginal likelihood of joint distribution over $\mathcal{G}$ is as follows

$$S(\mathcal{G}; D) = \sum_{i=1}^{q} S(X_i, PA_i^{\mathcal{G}}), \tag{18}$$

where $X_i$ is the variable in $\mathcal{G}$ with its parents $PA_i^{\mathcal{G}}$ in the graph.

### 4.3. Search Procedure

We use GES (Chickering, 2002) as the search algorithm with our proposed score function. Although recently there have been continuous optimization-based search methods (Zheng et al., 2018; 2020; Yu et al., 2019; Zhu et al., 2019), they cannot theoretically guarantee to find global minimizer. Moving forward, we will demonstrate that our methods are locally consistent. With this property, along with GES as the search method, we can ensure finding the Markov equivalence class consistent with the data generative distribution asymptotically. Regarding the locally consistent property, intuitively, a scoring criterion is locally consistent if adding or removing edges towards the correct causal relationship results in an increase in score, while doing the opposite leads to a decrease. More formally, the locally consistent property is defined as follows:

**Definition 4.1.** (Chickering, 2002, Locally Consistent Scoring Criterion, Definition 6) Let $D$ be the observation from some distribution $p(\cdot)$. Let $\mathcal{G}$ be any DAG, and let $\mathcal{G}'$ be the DAG that results from adding the edge $X_i \to X_j$. A scoring criterion $S(\mathcal{G}, D)$ is locally consistent if the following two properties hold as the sample size $n \to \infty$:

1. If $X_j \not\perp\!\!\!\perp_p X_i \mid PA_j^{\mathcal{G}}$, then $S(\mathcal{G}'; D) > S(\mathcal{G}; \mathcal{D})$.
2. If $X_j \perp\!\!\!\perp_p X_i \mid PA_j^{\mathcal{G}}$, then $S(\mathcal{G}'; D) < S(\mathcal{G}; D)$.

The following Lemma 4.2 shows that our score function is locally consistent.

**Lemma 4.2.** *Under the condition that,*

$$\lim_{n\to\infty} \frac{1}{3!}(\boldsymbol{\sigma} - \widehat{\boldsymbol{\sigma}})^3 \frac{\partial^3 \log p(X, PA^{\mathcal{G}} \mid \widehat{\boldsymbol{\sigma}})}{\partial \boldsymbol{\sigma}^3} = 0, \tag{19}$$

*our score function using the marginal likelihood of joint distributions is locally consistent.*

$\boldsymbol{\sigma}$ represents all the trainable parameters in the model, including the parameters in $k_{\mathcal{X}}$ and also that Gaussian process involved. And $\widehat{\boldsymbol{\sigma}}$ represents the true parameters. Lemma 4.2 illustrates that as the estimated parameters $\boldsymbol{\sigma}$ are close to the true parameters $\widehat{\boldsymbol{\sigma}}_\epsilon$ as $n \to \infty$, our score function is locally consistent. The proof is provided in Appendix A.4. With the local consistency property holding, using GES with our score function can guarantee finding the optimal Markov equivalence class (Chickering, 2002, Lemma 9).

### 4.4. Comparison with existing score functions

Although our approach and the method proposed by (Huang et al., 2018) are both rooted in RKHS regression framework, it is crucial to emphasize the significant differences in both the assumed relation model and the score function. The existing method assumes a regression model for the true relationship between the pre-fixed features of response variables, which necessitates heuristic selection of kernel parameters. While we considers it as a mixture of independent noise variables with learnable kernel functions, which is more flexible and capable of capturing more general causal relationships.

The differences in relationship modeling correspond to differences in the score function as well. While Huang et al. (2018) employs the conditional likelihood as the score function, it cannot be applied to our extended model due to the lacks of effective constraints on the parameters in $k_{\mathcal{X}}$, as illustrated in Figure 1(b). Therefore, we utilize mutual information-based score function, which is equivalent to the negative log-likelihood of the variables involved. Consequently, our score function can automatically learn the optimal kernel for the given data, performing superiorly compared to the median heuristic in most cases.

## 5. Experimental Results

We conducted experiments on both synthetic data and real-world benchmark datasets to evaluate the proposed score function. We compared our score function with the previous conditional likelihood-based score functions: (1) the vanilla conditional marginal likelihood conducted in the original data space (Friedman & Nachman, 2000), denoted as GP; (2) the RKHS-based conditional likelihood methods: CV (cross-validated likelihood) and Marg (marginal likelihood) (Huang et al., 2018). All the score-based methods were combined with GES (Chickering, 2002) as the search procedure. Additionally, we conducted a comparison with a constraint-based method PC, using PC (Spirtes et al., 2000) as the search algorithm with kernel-based KCI (Zhang et al., 2011) for conditional independent testing.

In all experiments involving kernel-based methods, we employed the widely-used Gaussian kernel. Consistent with the median heuristic setting, we set the kernel bandwidth as twice the median distance between the original input points for both PC, CV and Marg. Additionally, we initialized our method with this kernel bandwidth to demonstrate that our approach can begin with heuristic initialization and find improved kernel parameters through optimization. For more experimental details, please refer to Appendix B.3.2.

We use two accuracy measurement to evaluate their performance: (1) F1 score, which is a weighted average of the precision and recall, where higher F1 scores indicate higher accuracies; (2) normalized SHD score, the normalized structural hamming distance (Tsamardinos et al., 2006), which evaluates the difference between the recovered Markov equivalence class and the true class with correct directions. A lower SHD score indicates better accuracy.

Additionally, we also compared with recent popular con-

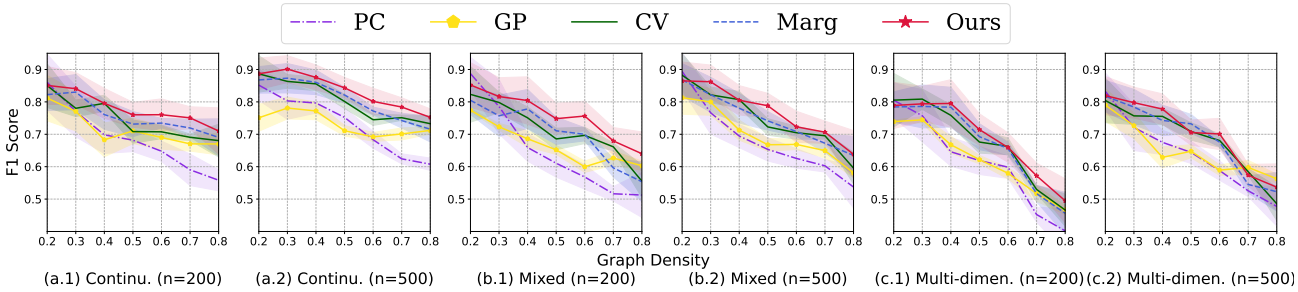



*Figure 2.* The F1 score of recovered causal graphs on: (a.1) Continuous data with sample size $n = 200$ and (a.2) $n = 500$; (b.1) mixed data with $n = 200$ and (b.2) $n = 500$; and (c.1) multi-dimensional data with $n = 200$ and (c.2) $n = 500$. The x-axis represents the graph density and the y-axis is the F1 score; higher F1 scores indicate higher accuracy. Shaded regions show standard errors for the mean.

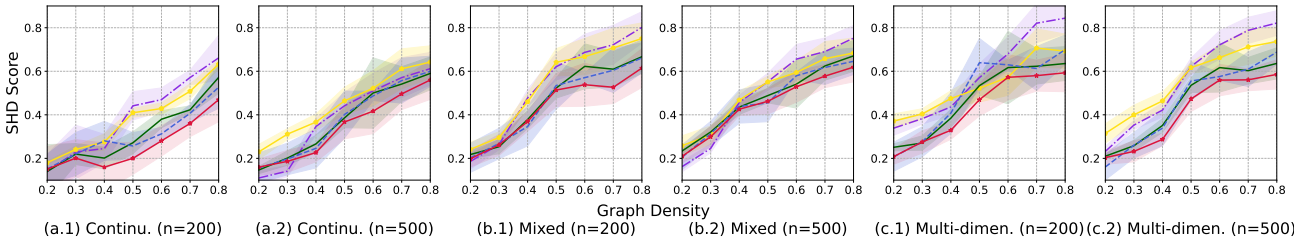



*Figure 3.* The normalized SHD of recovered causal graphs on synthetic data with different data types and sample sizes. The y-axis is the normalized SHD score and the lower SHD score means better accuracy.

tinuous optimization-based methods on real benchmarks. Please refer to Appendix B.4 for the experimental results and discussion.

### 5.1. Synthetic Data

We first conducted tests on synthetic data with various data types: (1) continuous data; (2) mixed data, comprising both continuous and discrete variables; and (3) multi-dimensional data, where variables may have more than one dimension. For each variable $X_i$ in the graph, the data was generated according to the following equation:

$$X_i = g_i(f_i(PA_i) + \varepsilon), \tag{20}$$

where $f_i$ and $g_i$ were randomly chosen from the *linear*, *sin*, *cos*, *tanh*, *exponential* and *power* functions. The noise term $\varepsilon$ was randomly chosen from either a *Gaussian* or *uniform* distribution. For the mixed data, we first generated continuous data and then randomly discretized some of the variables among them. For multi-dimensional data, we first transformed the parent variables to the same dimension as the response variable, and then the generation process was the same as for continuous variables. We generated causal graphs with varying graph densities ranging from 0.2 to 0.8. The graph density is measured by the ratio of the number of edges to the maximum possible number of edges in the graph. Each generated graph involves 8 variables, with sample sizes of $n = 200$ and 500. For each graph density, sample size and data type, we generated 20 realizations. For more details about the synthetic dataset, please refer to Appendix B.3.1.

Figure 2 presents F1 score of the recovered causal graphs using our proposed score function, along with compared methods. Generally, the accuracy increases along with the sample size and decreases along with the graph density across all methods. Some local fluctuations may arise due to the randomness in the generated relations. All methods perform worse on multi-dimensional data compared to other data types. The constraint-based PC demonstrates better performance when graphs are sparse but deteriorates as graph density increases. This can be attributed to the PC method relying on conditional independence tests, with the test power diminishing as the number of independent variables increases in dense graphs. GP, utilizing the vanilla marginal likelihood in the original data space, performs less effectively than the RKHS-based Marg and CV. This suggests that the RKHS regression model, with the feature projection on the response variables albeit fixed, better captures general relationships within the data compared to the vanilla regression model. Overall, our method exhibits the highest F1 score across most graph density settings. Particularly, when the graph density is relatively high, our approach outperforms conditional likelihood-based methods such as Marg and CV by a considerable margin.

Figure 3 displays the results of normalized SHD, which additionally considers recovered causal directions. A lower SHD signifies higher accuracy in the recovered graph. Over-

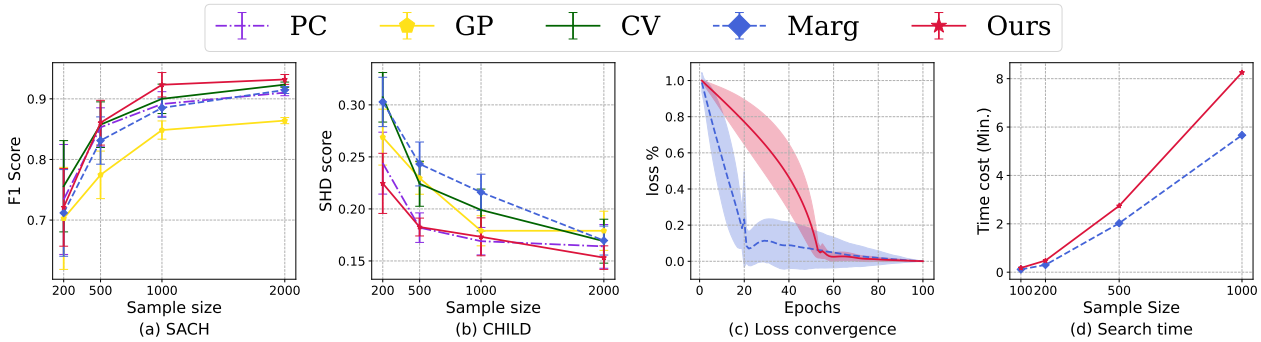



*Figure 4.* Results on benchmarks (a) SACH and (b) CHILD with different sample sizes. A higher F1 score or a lower SHD score indicates better performance. Comparison between Marg and our method on (c) convergence time for one single edge and (d) the overall search time for the entire graph.

all, we found that our kernel selection method gives the best accuracy in all cases, which is consistent with F1 score. In dense graph scenarios, our method exhibits much lower SHD scores compared to other methods. This suggests that our approach is more accurate in determining the correct direction, particularly when dealing with a larger number of variables. The experiment results underscore the effectiveness of our method, while also highlighting the crucial role of bandwidth selection. By modeling the joint distribution, our approach can identify optimal kernel parameters, surpassing the performance of median heuristic bandwidth choice. More experimental results on synthetic data can be found in Appendix B.1.

### 5.2. Real Benchmark Datasets

We further evaluated our method on two widely-used causal discovery benchmarks: SACH and CHILD networks. The SACH network comprises 11 variables and 17 edges, while the CHILD network consists of 20 variables with 25 edges. Therefore, both networks are relatively sparse in nature. The variables in both networks are discrete, with values ranging from 1 to 6. We randomly selected data with sample sizes of $n = 200, 500, 1000$ and $2000$, repeating 20 times for each sample size.

Figure 4(a) illustrates the F1 score of the recovered causal skeleton. Due to the sparsity of the networks, the constraint-based PC, which utilizes conditional independence tests, outperforms the score-based GP. However, as the sample size increases, our method gradually outperforms all others. We also employ SHD to evaluate the performance of all methods on the CHILD network, and the results are presented in Figure 4(b). From these results, it is evident that the SHD score of our method is significantly lower than that of Marg and CV. The experimental results on real benchmark datasets underscore the effectiveness of our method when applied to real-world data. They also suggest that median heuristically-selected kernels may not be universally suitable for various real-world datasets. The automatic selection of the optimal kernel in our method enables better handling of the wide range of causal relationships inherent in real-world scenarios. More experimental results on real data can be found in Appendix B.2 and Appendix B.4 (Comparison with continuous optimization-based methods).

### 5.3. Computation Analysis

We conducted a comparison of the computation cost between our method and Marg. To evaluate the convergence of our method and Marg, we randomly generated 100 sets of relationships and recorded the relative changes in loss during the training process. For the experimental details, please refer to Appendix B.3.4. As shown in Figure 4(c), our method, despite incorporating additional learnable parameters (parameters in $k_{\mathcal{X}}$), demonstrates a longer convergence time compared to Marg, yet achieves convergence at a relatively fast rate. We also compared the time required for searching the entire graph with various sample sizes. As depicted in Figure 4(d), the search time for the entire graph in our method is longer than that of Marg. However, considering that Gaussian Process poses a common computational bottleneck for both our method and Marg, with a complexity of $O(n^3)$, the additional time required by our method remains within acceptable limits.

## 6. Conclusion

In this paper, we presented a novel kernel selection method in the generalized score function for causal discovery, which allows for the automatic selection of the optimal kernel that best fits the data. By modeling the causal relation between variables as a mixture of independent noise variables, we utilize the marginal likelihood of the joint distribution as the score function to automatically select the optimal kernel specific to the given data. The experimental results on synthetic and real benchmarks demonstrated the effectiveness of our method compared to median heuristic-based approaches. In future research, we aim to enhance the computational

efficiency of our method and explore its integration with a broader range of search methods, including continuous optimization-based approaches, enabling our method to be applicable to datasets with greater number of variables.

## Acknowledgements

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. FL was supported by the Australian Research Council with grant numbers DP230101540 and DE240101089, and the NSF&CSIRO Responsible AI program with grant number 2303037. TLL was partially supported by the following Australian Research Council projects: FT220100318, DP220102121, LP220100527, LP220200949, and IC190100031. KZ would like to acknowledge the support from NSF Grant 2229881, the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Florin Court Capital. MG was supported by ARC DE210101624.

## Impact Statement

This paper presents work whose goal is to advance the field of Causal Discovery. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References

Adriaans, P. and van Benthem, J. *Handbook of Philosophy of Information*. Elsevier, 2008.

Antonakis, J. and Lalive, R. Counterfactuals and causal inference: Methods and principles for social research. *Structural Equation Modeling*, 18(1):152–159, 2011.

Baker, C. R. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186: 273–289, 1973.

Ben-Israel, A. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.

Bühlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Chapelle, O. and Vapnik, V. Model selection for support vector machines. *Advances in neural information processing systems*, 12, 1999.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3 (Nov):507–554, 2002.

Chu, C.-K. and Marron, J. S. Choosing a kernel regression estimator. *Statistical science*, pp. 404–419, 1991.

Friedman, N. and Nachman, I. Gaussian process networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 211–219, 2000.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5 (Jan):73–99, 2004.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Kernel dimension reduction in regression. *The Annals of Statistics*, 37 (4):1871–1905, 2009.

Garreau, D., Jitkrittum, W., and Kanagawa, M. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Geiger, D. and Heckerman, D. Learning gaussian networks. In *Proceedings of Conference on Uncertainty in Artificial Intelligence*, pp. 235–243, 1994.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schoelkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13, 2012a.

Gretton, A., Sriperumbudur, B., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., and Fukumizu, K. Optimal kernel choice for large-scale two-sample tests. In *Proceedings of Conference on Advances in Neural Information Processing Systems*, pp. 1205–1213, 2012b.

Haughton, D. M. On the choice of a model to fit data from an exponential family. *The annals of statistics*, 16:342–355, 1988.

Herrmann, E., Gasser, T., and Kneip, A. Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, 79(4):783–795, 1992.

Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. Generalized score functions for causal discovery. In *Proceedings of ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1551–1560, 2018.

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

Keropyan, G., Strieder, D., and Drton, M. Rank-based causal discovery for post-nonlinear models. In *International Conference on Artificial Intelligence and Statistics*, pp. 7849–7870. PMLR, 2023.

Kim, S.-J., Magnani, A., and Boyd, S. Optimal kernel selection in kernel fisher discriminant analysis. In *Proceedings of the 23rd international conference on Machine learning*, pp. 465–472, 2006.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pp. 6316–6326. PMLR, 2020.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR, 2016.

Londei, A., D 'Ausilio, A., Basso, D., and Belardinelli, M. O. A new method for detecting causality in fmri data of cognitive processing. *Cognitive processing*, 7:42–52, 2006.

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021.

Maxwell Chickering, D. and Heckerman, D. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine learning*, 29:181–212, 1997.

Moneta, A., Entner, D., Hoyer, P. O., and Coad, A. Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730, 2013.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Schölkopf, B., Smola, A. J., Bach, F., et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

Schwarz, G. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

Silander, T. and Myllymäki, P. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of Conference on Uncertainty in Artificial Intelligence*, pp. 445–452, 2006.

Sokolova, E., Groot, P., Claassen, T., and Heskes, T. Causal discovery from databases with discrete and continuous variables. *Lecture Notes in Computer Science; 8754*, pp. 442–457, 2014.

Spirtes, P. and Zhang, K. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pp. 1–28, 2016.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2021.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

Uemura, K., Takagi, T., Takayuki, K., Yoshida, H., and Shimizu, S. A multivariate causal discovery based on post-nonlinear model. In *Proceedings of Conference on Causal Learning and Reasoning*, pp. 826–839, 2022.

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

Williams, C. and Rasmussen, C. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019.

Yuan, C. and Malone, B. Learning optimal bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48:23–65, 2013.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of Conference on Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.

Zhang, K., Schölkopf, B., Spirtes, P., and Glymour, C. Learning causality and causality-related learning: some recent progress. *National science review*, 5(1):26–29, 2018.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

Zhu, S., Ng, I., and Chen, Z. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2019.

In this section, we present all the proofs and more experimental details and results. The content of the Appendix is as follows:

- **Proofs**
  - A.1 Connection between mutual information and joint likelihood.
  - A.2 Derivation of the Jacobian determinant.
  - A.3 Derivation of marginal likelihood of joint distribution in RKHS.
  - A.4 Proof of Lemma 4.2.
- **More experimental details and results**
  - B.1 More experiment result on synthetic data.
  - B.2 More experiment result on real benchmarks.
  - B.3 Experiment implementation details
  - B.4 Comparison with continuous optimization-based methods.

# A. Proofs

## A.1. Connection between mutual information and joint likelihood

Suppose there are $n$ observation $(\boldsymbol{x}, \boldsymbol{pa}) = \left\{(x^1, pa^1), (x^2, pa^2), \cdots, (x^n, pa^n)\right\}$. According to Eq. 8, for one particular observation $(x, pa)$, the following relation holds

$$\boldsymbol{k}_x = f(pa) + \boldsymbol{\varepsilon}, \tag{21}$$

where $\boldsymbol{k}_x = [k_{\mathcal{X}}(x_1, x), k_{\mathcal{X}}(x_2, x), \cdots, k_{\mathcal{X}}(x_n, x)]^T$ is the empirical feature vector of $x$, $f : \mathcal{P} \to \mathcal{F}_{\mathcal{X}}$ is a nonlinear function, and $\boldsymbol{\varepsilon}$ represents the independent noise vector. For each dimension in $\boldsymbol{k}_x$, we denote $g_i(\cdot) = k_{\mathcal{X}}(x_i, \cdot)$ and $g_i$ are determined by the parameters in $k_{\mathcal{X}}$. If we further assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2 I)$, then the relation can be formulated as

$$g_i(x) = f(pa) + \varepsilon_i, i = 1, 2, \ldots, \tag{22}$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. And the likelihood can be represented as

$$p(g_i(x), pa \mid f, g_i) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(g_i(x) - f(pa))^2}{2\sigma_\varepsilon^2}\right) \tag{23}$$

Based on the change-of-variables theorem, we have

$$p(PA, X) = p(PA, g_i(X)) \cdot |\det \mathbf{J}_i|, \tag{24}$$

with the Jacobian $\mathbf{J}_i = [\partial(PA, g_i(X))/\partial(PA, X)]$. Since $\partial\tilde{\varepsilon}_i/\partial x = \partial g_i(X)/\partial x$ holds in our model, the Jabobian in Eq. 24 is consistent with Eq. 14. Therefore, the likelihood of joint distribution $p(X, PA)$ under the model in Eq. 22 is

$$\begin{aligned} L(f, g_i) &= \sum_j \log p(x^j, pa^j \mid f, g_i) \\ &= \sum_j \log p(g_i(x^j), pa^j \mid f, g_i) + \sum_j \log|\det \mathbf{J}_i^j| \\ &= \sum_j \log p(\tilde{\varepsilon}_i^j) + \sum_j \log|\det \mathbf{J}_i^j|, \end{aligned} \tag{25}$$

where $\tilde{\varepsilon}_i^j = g_i(x^j) - f(pa^j)$. Integrating the each dimension of the estimated noise, the log likelihood of joint distribution $p(X, PA)$, the likelihood of $p(X, PA)$ under the model in Eq. 21 is

$$L(f, k_{\mathcal{X}}) = \sum_i \sum_j \log p(\tilde{\varepsilon}_i^j) + \sum_i \sum_j \log|\det \mathbf{J}_i^j|, \tag{26}$$

which is consistent with the first two term in Eq. 12. Therefore, minimizing the mutual information is equivalent to maximize the likelihood of joint distribution $p(X, PA)$ in our extended model.

### A.2. Derivation of the Jacobian determinant

Suupose that the random variable $X$ and $PA$ have $l$ and $m$ dimension respectively and $(x^j, pa^j) \in (\mathcal{X}, \mathcal{P})$ is one of the $n$ observations in $(\boldsymbol{x}, \boldsymbol{pa})$. We represent $x^j = [x_1^j, \cdots, x_l^j]^\top$ and $pa^j = [pa_1^j, \cdots, pa_m^j]^\top$. Then the Jacobian $\mathbf{J}$ has the following form

$$\mathbf{J} = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \hline -f'(pa_1^j) & \cdots & -f'(pa_m^j) & K'_{X(j,i)}(1) & \cdots & K'_{X(j,i)}(l) \end{array}\right], \tag{27}$$

which is a $(m+1) \times (m+l)$ volume with $f'(pa_i^j) = \partial f(pa^j)/\partial pa_i^j$ and $\mathbf{K}'_{X(j,i)}(t) = \partial k_{\mathcal{X}}(x^j, x^i)/\partial x_t^j$, and here we denote them as $f_i'$ and $K_t'$ for simplicity. According to (Ben-Israel, 1999), the Jacobian determinant can be replaced by the volume of the Jacobian that

$$|\det \mathbf{J}| = \mathrm{vol}\mathbf{J} = \sqrt{\det \mathbf{J}\mathbf{J}^\top} = \det \left[\begin{array}{ccc|c} 1 & 0 & 0 & -f_1' \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & 1 & -f_m' \\ \hline -f_1' & \cdots & -f_m' & \sum \end{array}\right]^{\frac{1}{2}}, \tag{28}$$

where $\sum = \sum_i f_i'^2 + \sum_t K_t'^2$. Based on the properties of determinants, we have

$$|\det \mathbf{J}| = \left[\begin{array}{ccc|c} 1 & 0 & 0 & -f_1' \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & 1 & -f_m' \\ \hline 1 & \cdots & 1 & \sum - \sum_i f_i'^2 \end{array}\right]^{\frac{1}{2}} = \sqrt{\sum_t K_t'^2} \tag{29}$$

Therefore, we can derive Eq. 16

### A.3. Derivation of marginal likelihood of joint distribution in RKHS

Suppose there are $n$ observation $(\boldsymbol{x}, \boldsymbol{pa}) = \{(x^1, pa^1), (x^2, pa^2), \cdots, (x^n, pa^n)\}$ and we use Gaussian process to model the nonlinear function $f$, which is $f \sim \mathcal{N}(0, K_{PA})$ with $K_{PA}$ the covariance matrix of $PA$ and $K_{PA(ij)} = k_{\mathcal{P}}(pa^i, pa^j)$. We denote $\sigma_p$ and $\sigma_x$ as the trainable parameters in $k_{\mathcal{P}}$ and $k_{\mathcal{X}}$ respectively. Therefore, according to Eq. 25, for one particular observation $(x, pa)$, the log likelihood of $p(x, pa)$ is represented as

$$\begin{aligned}
\log p(X = x, PA = pa \mid \sigma_x, \sigma_p) &= \log \int p(x, pa \mid f, \sigma_x) \cdot p(f \mid \sigma_p)\, \mathrm{d}f \\
&= \log \int p(\boldsymbol{k}_x, pa \mid f, \sigma_x) \cdot p(f \mid \sigma_p)\, \mathrm{d}f + \sum_{i=1}^{n} \log \int |\det \mathbf{J}_i| \cdot p(f \mid \sigma_p)\, \mathrm{d}f \\
&= \log \int p(\boldsymbol{k}_x, pa \mid f, \sigma_x) \cdot p(f \mid \sigma_p)\, \mathrm{d}f + \sum_{i=1}^{n} \log |\det \mathbf{J}_i| \\
&= \log \int p(\tilde{\varepsilon}) \cdot p(f \mid \sigma_p)\, \mathrm{d}f + \sum_{i=1}^{n} \log |\det \mathbf{J}_i| \\
&= \log \int \mathcal{N}(0, \sigma_\varepsilon^2 I) \cdot \mathcal{N}(0, K_{PA})\, \mathrm{d}f + \sum_{i=1}^{n} \log |\det \mathbf{J}_i| \\
&= \log\, \mathcal{N}(0, K_{PA} + \sigma_\varepsilon^2 I) + \sum_{i=1}^{n} \log |\det \mathbf{J}_i|
\end{aligned}$$

where $\sigma_p$ is the parameters in $K_{PA}$, $\sigma_x$ is the parameters of $l_{\mathcal{X}}$ and $\sigma_\varepsilon$ is the regularization parameter and all of them are trainable in our model. Let us further denote $K_\theta = (K_{PA} + \sigma_\varepsilon^2 I)$. Therefore, the log likelihood of joint density $(X, PA)$ on

the observations $(\boldsymbol{x}, \boldsymbol{pa})$ is represented as

$$\begin{aligned}
L(\sigma_x, \sigma_p, \sigma_\varepsilon) &= \prod_{j=1}^{n} \left\{ \log \mathcal{N}(0, K_\theta) + \sum_{i=1}^{n} \log |\det \mathbf{J}_i^j| \right\} \\
&= -\frac{1}{2} \sum_{j=1}^{n} \boldsymbol{k}_{x(j)}^T K_\theta^{-1} \boldsymbol{k}_{x(j)} - \frac{n}{2} \log \det |K_\theta| - \frac{n^2}{2} \log 2\pi + \sum_j \sum_i \log |\det \mathbf{J}_i^j| \\
&= -\frac{1}{2} \sum_{j=1}^{n} (K_X K_\theta^{-1} K_X)_{jj} - \frac{n}{2} \log \det |K_\theta| - \frac{n^2}{2} \log 2\pi + \sum_{i \neq j}^{n} \log \sqrt{\sum_t K'^2_{X(j,i)}(t)} \\
&= -\frac{1}{2} \mathrm{trace}(K_X K_\theta^{-1} K_X) - \frac{n}{2} \log \det |K_\theta| - \frac{n^2}{2} \log 2\pi + \sum_{i \neq j}^{n} \log \sqrt{\sum_t K'^2_{X(j,i)}(t)}
\end{aligned}$$

where $j$ represents the $j$-th sample in the observation $(\boldsymbol{x}, \boldsymbol{pa})$, $i$ represents the $i$-th dimension in the vector and $K_X$ and $K_{PA}$ are the kernel matrices of $X$ and $PA$ with $K_{X(ij)} = k_{\mathcal{X}}(x^i, x^j)$ and $K_{PA(ij)} = k_{\mathcal{P}}(pa^i, pa^j)$.

### A.4. Proof of Lemma 4.2

Our score function using marginal likelihood of joint density $p(X, PA^{\mathcal{G}})$ is

$$S(X, PA^{\mathcal{G}}) = \log p(X, PA^{\mathcal{G}} | \boldsymbol{\sigma}), \tag{30}$$

And $\boldsymbol{\sigma}$ represents all the trainable parameters in the model with $\boldsymbol{\sigma} = [\sigma_x, \sigma_p, \sigma_\varepsilon]$ in our extended model. $\sigma_x$ represents the parameters in $k_{\mathcal{X}}$ while $\sigma_p$ and $\sigma_\varepsilon$ are the parameters involved in the Gaussian process. We define $\widehat{\boldsymbol{\sigma}} = [\widehat{\sigma}_x, \widehat{\sigma}_p, \widehat{\sigma}_\varepsilon]$ to be the true parameters with the maximum likelihood of $p(X, PA^{\mathcal{G}})$. By employing the Laplace method (Maxwell Chickering & Heckerman, 1997), we can derive that

$$\log p(X, PA^{\mathcal{G}} | \boldsymbol{\sigma}) \approx \log p(X, PA^{\mathcal{G}} \mid \widehat{\boldsymbol{\sigma}}) + \log p(\widehat{\boldsymbol{\sigma}}) + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log |A|, \tag{31}$$

where $d$ is the number of trainable parameters in the model ($d = 3$ in our method) and $A$ is the negative Hessian of $\log p(X, PA^{\mathcal{G}} | \boldsymbol{\sigma})$ evaluated at $\widehat{\boldsymbol{\sigma}}$. The first term, $\log p(X, PA^{\mathcal{G}} | \widehat{\boldsymbol{\sigma}})$, increases linearly with the sample size $n$. And $\log |A|$ increases as $d \log n$. The remaining two terms $\log p(\widehat{\boldsymbol{\sigma}})$ and $\frac{d}{2} \log(2\pi)$ are both consistent with $n$ increasing. That is, for a large sample size $n$, we obtain

$$\log p(X, PA^{\mathcal{G}} | \boldsymbol{\sigma}) \approx \log p(X, PA^{\mathcal{G}} \mid \widehat{\boldsymbol{\sigma}}) - \frac{d}{2} \log n. \tag{32}$$

This approximation is equivalent to the Bayesian Information Criterion (BIC) (Schwarz, 1978). Since BIC is consistent (Haughton, 1988), the marginal likelihood of joint distribution $p(X, PA^{\mathcal{G}})$ as score function is also consistent.

Consequently, we have demonstrated the local consistency of our proposed score under the conditions in Lemma 4.2. Consider a given variable $X$ and all its potential parent variables combinations, denoted as $[PA_1, PA_2, \cdots, PA_n]$. Each combination $PA_i$ with $X$ corresponds to its respective maximum marginal likelihood of $p(X, PA_i \mid \widehat{\boldsymbol{\sigma}})$, denoted as $S_i$. The local consistency property guarantees that among all these combinations, the score for the correct parent set $PA_{true}$ for $X$, denoted as $S_{true}$, is the highest. Hence, our proposed score function can effectively identify the correct $PA$ variables of $X$ from among all potential parent sets.

## B. More experimental details and results

In this subsection, we provide more experimental results on the synthetic data (Appendix B.1) and real benchmarks (Appendix B.2). We also present the implementation details of the synthetic dataset and the compared methods. (Appendix B.3). In Appendix B.4, we experimentally compared our method with the continuous optimization-based methods.

### B.1. More experiment result on synthetic data

We present the experiment result on the synthetic discrete dataset, where all the variables are discrete. The details of the experiment are consistent with those on mixed data, with the discrete ratio $r$ set to 1. The results are shown in Figure 5. Figure 6 shows the results of our method and the comparison methods on continuous, mixed and multi-dimensional synthetic data with a sample size $n = 1000$.

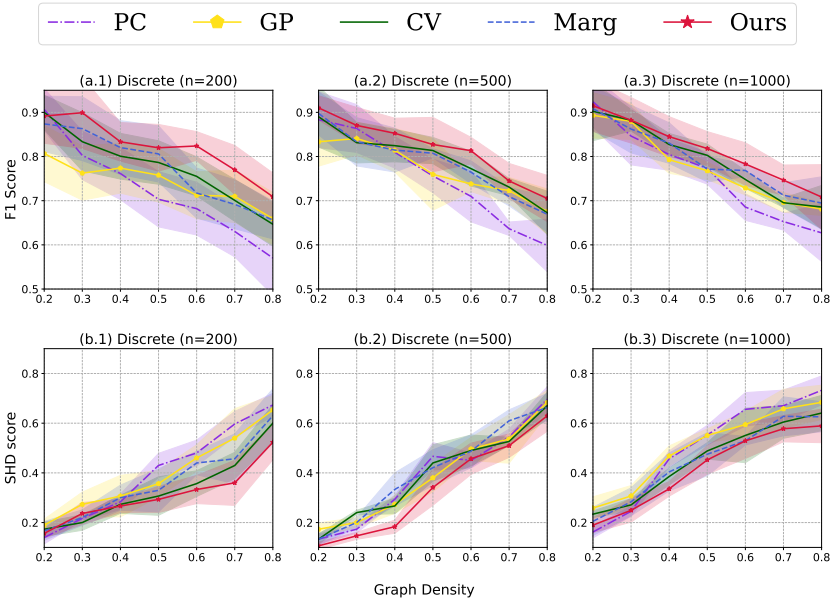



*Figure 5.* Results on synthetic discrete dataset. All the variables are discrete in the graph, with the value either from $[1, 5]$ or $[1, 20]$. The F1 score of recovered causal graphs with sample size (a.1) $n = 200$ and (a.2) $n = 500$ and (a.3) $n = 1000$, where a higher F1 score ↑ indicates greater accuracy. The normalized SHD score ↓ with different samples are presented in (b.1) $n = 200$ and (b.2) $n = 500$ and (b.3) $n = 1000$, with a lower SHD score signifying better accuracy. The x-axis is the graph density. Shaded regions show standard errors for the mean.

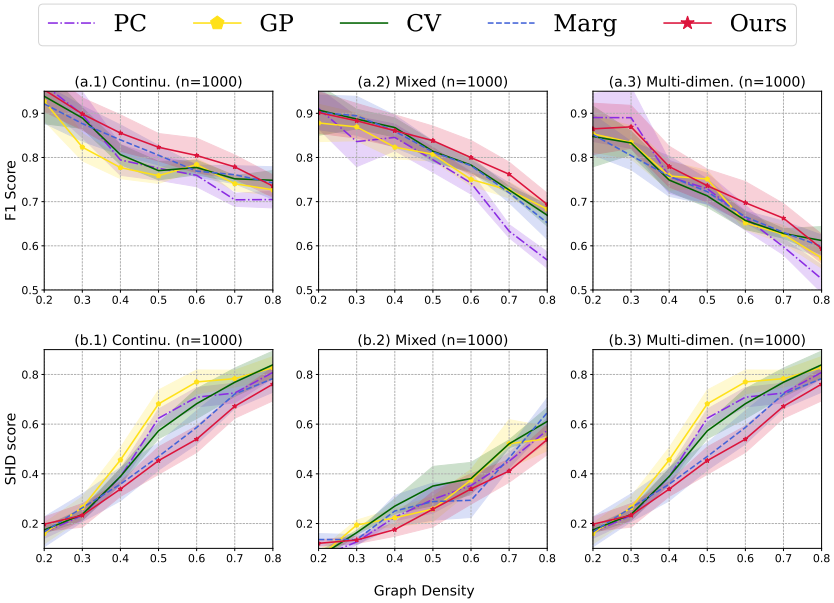



*Figure 6.* Results on synthetic dataset with sample size $n = 1000$. The F1 score ↑ of recovered graphs on: (a.1) continuous dataset; (a.2) mixed dataset; and (a.3) multi-dimensional dataset. The normalized SHD score ↓ of recovered graphs on: (b.1) continuous dataset; (b.2) mixed dataset; and (b.3) multi-dimensional dataset. The x-axis is the graph density.

### B.2. More experiment result on real benchmarks

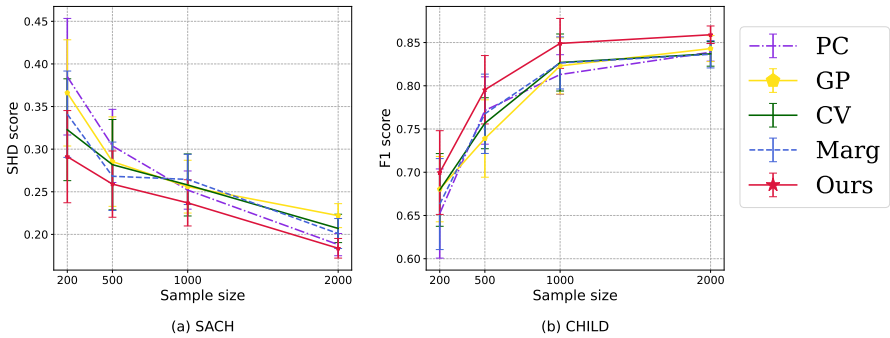



*Figure 7.* Results on benchmarks (a) SACH and (b) CHILD with different sample sizes.

### B.3. Experimental details

In this section, we provide the synthetic dataset details (Appendix B.3.1) and implementation details of our method and compared methods (Appendix B.3.2). We also offer more experiment details in Motivation (Appendix B.3.3) and Computation Analysis (Appendix B.3.4).

#### B.3.1. Synthetic Dataset

In the experiment on synthetic data, the generated graph consists of a total of 8 variables. Based on the given graph density, we randomly generate the corresponding number of edges in the graph. The leaf node $X_{\text{leaf}}$ in the graph, which has an in-degree of 0, follows either a normal distribution $\mathcal{N}(0,1)$ or a uniform distribution $U(-1,1)$ with equal probability. For each response variable $X_i$ in the graph, its data is generated according to the following causal relation:

$$X_i = g_i(f_i(PA_i) + \varepsilon), \tag{33}$$

where $PA_i$ represents its parent nodes in the graph. $f_i$ is equally likely to be sampled from *linear, sin, cos, tanh, exp* and $x^\alpha$. The *linear* function has two weight options: $0.5$ and $2.5$, and $\alpha$ in $x^\alpha$ is randomly selected from $1, 2, 3$. $\varepsilon$ is the additive noise, randomly following either $\mathcal{N}(0, 0.5)$ or a uniform distribution $U(-0.5, 0.5)$ with equal probability.

For the mixed data, which contains both discrete and continuous variables, we first generate the continuous data using the aforementioned process. Then, we randomly select variables according to the specified ratio $r$ and discretize them, with the value range either from $[1, 5]$ or $[1, 20]$. In our experiments, the ratio $r$ is set to $0.5$ for mixed data and $r = 1$ for discrete data, and its results have been shown in Appendix B.1

For multi-dimensional variables, we randomly assign dimensions to all variables, with dimensions ranging from 1 to 5. For leaf nodes, its data are generated in the same manner. As for the response variable $X$ and its parent node $PA$, we first multiply $PA$ with a matrix with all elements equal to 1, transforming its dimension to match that of $X$. Subsequently, we use Equation (33) to obtain the corresponding dependent variable.

#### B.3.2. Implementation Details

We present the implementation details for the learning procedure of our methods. Our model involves three parameters: (1) $\sigma_x$, the bandwidth of the kernel applied to the response variable; (2) $\sigma_p$, the parameter for the kernel function of the dependent variable; and (3) $\sigma_\varepsilon$, the regularization parameter in the inverse matrix in Equation (17). All these parameters are learnable in our method, eliminating the need for hyperparameter tuning. To avoid numerical issues, we explicitly set the ranges for these parameters as $\sigma_x, \sigma_p \in [0.1, 10]$ and $\sigma_\varepsilon \in [0.001, 10]$. We employed the L-BFGS (Liu & Nocedal, 1989) as the optimization method for our model with the default hyper-parameter setting[1]. We use GES (Chickering, 2002) as our search algorithm, and its implementation is highly-based on *causal-learn* package[2]. Our code is implemented in Python 3.8.7 and PyTorch 1.11.0.

[1] https://gist.github.com/arthurmensch/c55ac413868550f89225a0b9212aa4cd
[2] https://github.com/py-why/causal-learn

**Baseline.** We use most existing implementations for the compared methods.

- **PC** (Spirtes et al., 2000) is a widely-used search algorithm. And we use the kernel-based conditional independence, KCI (Zhang et al., 2011), to test conditional independence relationships. The KCI test can handle nonlinear causal relations and data from arbitrary distributions. The implementations of PC and KCI are based on *causal-learn* package. We used the default parameter settings, with a significance level of $\alpha = 0.05$. The median heuristic was also employed to initialize the kernel bandwidth involved.
- **GP** (Friedman & Nachman, 2000) is based on the regression model in the original data space, $X = f(PA) + \varepsilon$, and $f$ is modeled via the Gaussian process. We use the original marginal likelihood of the conditional distribution as its score function. We also used the median heuristic to initialize the kernel bandwidth on the parent nodes $PA$.
- **CV** (Huang et al., 2018) is based on the RKHS regression model and uses cross-validation to avoid overfitting of $f$. Its code was provided in *causal-learn* package. We used the default setting with 10-fold cross validation and the regularization parameter $\lambda = 0.01$
- **Marg** (Huang et al., 2018) is also based on the RKHS regression model. And it employ the marginal likelihood to avoid overfitting problem. For a fair comparison, we carefully implemented the code by ourselves. Apart from the trainable parameters in $k_{\mathcal{X}}$ and the score function, the rest of the implementation remains consistent with our method.

For a fair comparison, all methods are experimented on the same data instances and hardware environments.

#### B.3.3. More experiment details in Motivation

In Section 3, through a toy experiment, we show that the median heuristic may not effectively fit the given data, consequently failing to identify the correct relationships in the provided example. In this toy experiment, we generated a total of $n = 500$ samples. Hence, for each pair of observation $(x, y, z)$ in $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = [(x^1, y^1, z^1), \cdots, (x^n, y^n, z^n)]$, it is associated with a empirical feature vector $\boldsymbol{k}_x = \left[k_{\mathcal{X}}(x, x^1), \cdots, k_{\mathcal{X}}(x, x^n)\right]^\top$, where $k_{\mathcal{X}}$ is the kernel function. We also obtained the estimated noise vector $\tilde{\boldsymbol{\varepsilon}}_{X|Y} = \boldsymbol{k}_x - f(y)$. Both $\boldsymbol{k}_x$ and $\tilde{\boldsymbol{\varepsilon}}_{X|Y}$ have $n$-dimensions, and we denote $\tilde{\boldsymbol{\varepsilon}}_{X|Y} = [\tilde{\varepsilon}_1, \cdots, \tilde{\varepsilon}_n]$. We visualize the first dimension of the estimated noise $\tilde{\varepsilon}_1$, along with $Z$. The estimated noise $\tilde{\varepsilon}_1$ in Figure 1(c) are learned under a fixed $k_{\mathcal{X}}$ with the median heuristically-selected bandwidth. While Figure 1(d) represents the results obtained using our score function with trainable $k_{\mathcal{X}}$.

To further quantify the dependence between the estimated noise $\tilde{\boldsymbol{\varepsilon}}_{X|Y}$ and $Z$, we use the Hilbert-Schmidt independence criterion (HSIC) to measure their independence (Gretton et al., 2005). Suppose $X$ and $Y$ are two random variables on domains $\mathcal{X}$ and $\mathcal{Y}$ with the sample $(x, y)$, the HSIC of $X$ and $Y$ can be expressed in terms of kernels (Gretton et al., 2007):

$$\begin{aligned}\mathrm{HSIC}(X,Y) =& E_{XY}E_{X'Y'}[k_{\mathcal{X}}(x,x')k_{\mathcal{Y}}(y,y')] \\ &+ E_X E_{X'}[k_{\mathcal{X}}(x,x')] \cdot E_Y E_{Y'}[k_{\mathcal{Y}}(y,y')] \\ &- 2E_{X'Y'}[E_X k_{\mathcal{X}}(x,x') \cdot E_Y k_{\mathcal{Y}}(y,y')],\end{aligned} \tag{34}$$

where $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are two kernel functions. HSIC can serve as a measure of dependence between $X$ and $Y$. A high HSIC value suggests a strong statistical dependence between $X$ and $Y$, while a low HSIC value implies a tendency towards independence. We use a biased empirical estimate of HSIC to measure the independence between $\tilde{\varepsilon}_1$ and $Z$, which has the following form (Gretton et al., 2007):

$$\mathrm{HSIC}_b(\tilde{\boldsymbol{\varepsilon}}_1, \boldsymbol{z}) = \frac{1}{n^2}\mathrm{Tr}(K_\varepsilon H K_Z H), \tag{35}$$

where $n$ is the sample size, $K_\varepsilon$ and $K_Z$ are the kernel matrices with entries $K_{\varepsilon(i,j)} = k_\varepsilon(\tilde{\varepsilon}_{1(i)}, \tilde{\varepsilon}_{1(j)})$ and $K_Z(i,j) = k_{\mathcal{Z}}(z_i, z_j)$. $\tilde{\varepsilon}_{1(i)}$ is the first dimension of the estimated noise from $i$-th observation $(y_i, x_i)$ with the corresponding $z_i$. $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ with $I$ and $\mathbf{1}$ being the $n \times n$ identity matrix and the vector of 1's, respectively. For the kernel function $k_\varepsilon$ and $k_{\mathcal{Z}}$ in Eq. 35, we used a Gaussian kernel with kernel bandwidth set to the median distance between points in input space.

Since the dimension of $\tilde{\varepsilon}_{X|Y}$ is high, we choose to independently compute the HSIC between $Z$ and each dimension $\tilde{\varepsilon}_i$ in $\tilde{\varepsilon}_{X|Y}$ and then average them to obtain the final result. As shown in Figure 1(c) and (d), the HSIC value of our method (0.0038) are smaller than the median heuristic-based one's (0.0062). It indicates that the estimated noise $\tilde{\varepsilon}_{X|Y}$ in our method is more independent from $Z$, meaning our method is better at blocking the influence of $Z$ on $X$.

#### B.3.4. More experiment details in Computation Analysis

To compare the convergence of our method, we generated 100 random relationships examples according to Eq. 20 with sample size $n = 200$. All variables in the generated relationships are continuous and only have one dimension with at most 3 parent variables. As the L-BFGS algorithm automatically stops based on the gradient transformation during the learning process, we used Adam optimizer (Kingma & Ba, 2014), with an initial learning rate of 0.005 and a maximum epoch of 100. The losses of each relationship were normalized to reflect the relative changes in the score during the training process. In the experiment comparing the searching time for the entire graph, we generated the data with 8 continuous variables and the graph density of 0.5. For each sample size, we repeated 5 times.

### B.4. Comparison with continuous optimization-based approaches

Recently, there has been an emergence of continuous optimization-based approaches for causal discovery (Zheng et al., 2018; 2020; Yu et al., 2019; Zhu et al., 2019; Lorch et al., 2021). These methods perform continuous optimization on an adjacency matrix by employing functions to enforce acyclicity, aiming to find the adjacency matrix with the minimum of some criterion. Based on this, they recast the graph-search problem into a continuous optimization problem, using neural networks and gradient descent to search for the complete DAGs. In contrast, the traditional discrete search-based approaches, such as GES (Chickering, 2002) and PC (Spirtes et al., 2000), perform local searches on subsets of variables to compose the complete causal graph.

**Difference.** We want to emphasize the difference in the starting point between such continuous optimization-based approaches and our proposed score function. The continuous optimization-based approaches assume a relatively simple local relationship, such as a linear model or nonlinear additive model. Under the assumption of a relatively simple local relationship, such methods focus on searching for the entire graph while adhering to the acyclicity constraint. Its advantage lies in avoiding the super-exponentially growing potential DAG candidates with the number of variables that traditional discrete search methods entail. Our approach, on the other hand, focuses on modeling more generalized local relationships compared to the linear or nonlinear additive models used by the former. In our approach, we extend the nonlinear additive model by introducing a trainable kernel function applied to the response variable, which is general. Correspondingly, our score function utilizing the discrete search method GES requires much longer search times compared to continuous optimization-based methods and cannot be applied to networks with a large number of variables. Next, we compared our method with these continuous optimization methods on the real benchmarks SACHS and CHILD.

**Implementations.** The selected continuous optimization-based methods for comparison are: NOTEAR (Zheng et al., 2018), NS-MLP (Zheng et al., 2020), DAG-GNN (Yu et al., 2019) and Dibs (Lorch et al., 2021). We used their default parameter settings:

- **NOTEARS**[2] : We adopt the same hyper-parameters suggested in the original papers with the graph threshold set to 0.3, $h_{min} = 1 \times 10^{-8}$ and $\rho_{max} = 1 \times 10^{16}$. We used $l_2$ loss with sparse parameter $\lambda_1 = 0.1$.
- **NS-MLP**[2] : the $l_2$ penalty to the multi-layer perceptrons $\lambda_2$ is set to 0.01 with $\lambda_1 = 0.01$. The remaining parameters are set the same with NOTEARS.
- **DAG-GNN**[3]: We utilized a Multilayer Perceptron (MLP) for both the encoder and decoder, with the penalty parameter's initial value $c = 1.0$, Lagrange multiplier $\lambda_A = 0$ and $\tau_A = 0$. The graph threshold is set as 0.3 as suggested.
- **Dibs**[4]: We used nonlinear Gaussian model with $\alpha = 0.05$, $\beta = 1.0$ and the bandwidth $\gamma_z = 5$, $\gamma_\theta = 500$. We generated 10 particles and selected the one with the highest F1 score as the final recovered graph.

**Discussion.** The results are shown in Figure 8. Notice that the true graph is indeed sparse and an empty graph can have a relative low SHD score. Therefore, we use more specific metrics: precision (left), recall (center), and their combined F1 score[5] (right) to evaluate them. From the results, we can see that most continuous optimization methods have high precision, except for Dibs. However, the recall of these optimization-based methods is significantly lower than that of our

[2] https://github.com/xunzheng/notears

[3] https://github.com/fishmoon1234/DAG-GNN

[4] https://github.com/larslorch/dibs

[5] F1 score is a weighted average of the precision and recall, with $F1 = \frac{2\,\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}}$.

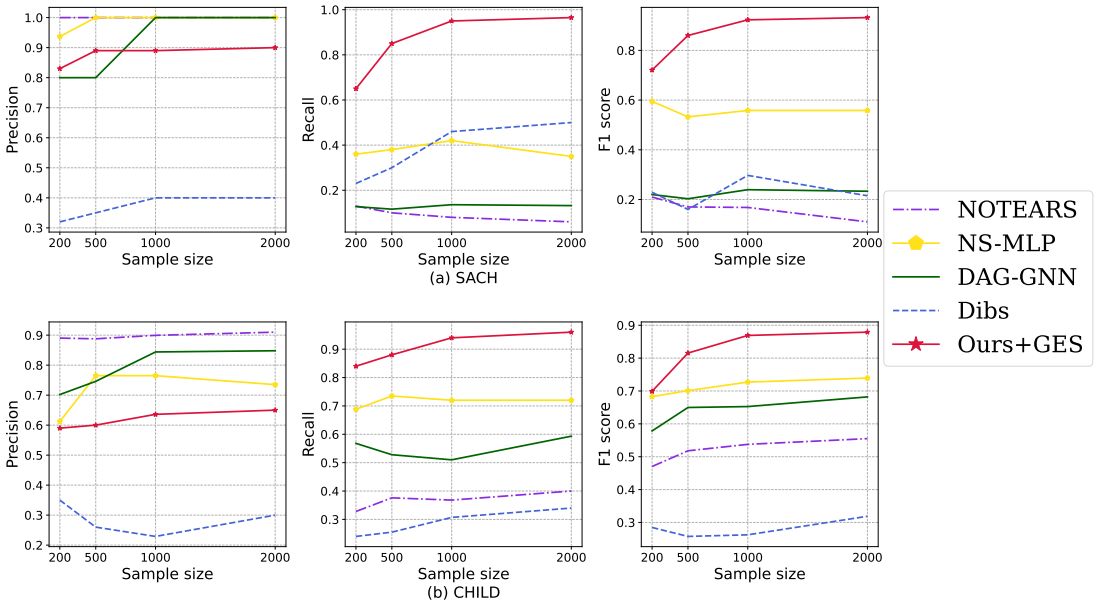



*Figure 8.* Comparison with optimization-based methods on real benchmark (a) SACH and (b) CHILD with different sample sizes. The x-axis is the sample size and the y-axis is precision (left), recall (center) and F1 score (right).

discrete search-based method, indicating that these methods miss many edges during the search process. This ultimately leads to the poor performance of these optimization-based methods in terms of F1 score evaluation, as shown in the right subplots of Figure 8. What we want to emphasize is that the significant performance gap in F1 score between our method and continuous optimization-based methods is mainly due to the difference in search strategy. Discrete search methods, which fully search over the possible combinatorial space, are therefore more accurate but also require more search time. In the future, we are exploring incorporating our proposed score function into these continuous optimization-based methods to make our score function applicable to the datasets with a larger number of variables.