Generalized Score Functions for Causal Discovery

Biwei Huang
Carnegie Mellon University
biweih@andrew.cmu.edu

Kun Zhang
Carnegie Mellon University
kunz1@andrew.cmu.edu

Yizhu Lin
Carnegie Mellon University
yizhul@andrew.cmu.edu

Bernhard Schölkopf
MPI for Intelligent Systems
bs@tuebingen.mpg.de

Clark Glymour
Carnegie Mellon University
cg09@andrew.cmu.edu

ABSTRACT

Discovery of causal relationships from observational data is a fundamental problem. Roughly speaking, there are two types of methods for causal discovery, constraint-based ones and score-based ones. Score-based methods avoid the multiple testing problem and enjoy certain advantages compared to constraint-based ones. However, most of them need strong assumptions on the functional forms of causal mechanisms, as well as on data distributions, which limit their applicability. In practice the precise information of the underlying model class is usually unknown. If the above assumptions are violated, both spurious and missing edges may result. In this paper, we introduce generalized score functions for causal discovery based on the characterization of general (conditional) independence relationships between random variables, without assuming particular model classes. In particular, we exploit regression in RKHS to capture the dependence in a nonparametric way. The resulting causal discovery approach produces asymptotically correct results in rather general cases, which may have nonlinear causal mechanisms, a wide class of data distributions, mixed continuous and discrete data, and multidimensional variables. Experimental results on both synthetic and real-world data demonstrate the efficacy of our proposed approach.

ACM Reference Format:
Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. 2018. Generalized Score Functions for Causal Discovery. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220104

1 INTRODUCTION

Traditionally, interventions or randomized experiments are used for inferring causal relationships. However, conducting such experiments is often expensive or even impossible, and from the results it is not easy to construct quantitative causal models. Alternatively, one may perform causal discovery from passively observational data, which has been made possible under proper assumptions [22, 28]. The approaches to causal discovery from observational data proposed over the past decades roughly fall into two categories, namely, constraint-based methods and score-based methods. For surveys of some of the recently proposed methods, one may refer to [29] and [36].

Constraint-based methods use statistical tests (conditional independence tests) to find the causal skeleton and determine the orientations of the edges up to the Markov equivalence class; all members of such a class have the same conditional independence relationships. In principle, constraint-based methods do not assume any particular form of causal mechanisms, given that the conditional independence test is reliable. As a consequence, they can be easily extended to handle more complex situations, such as the case with nonstationary time series or multiple, heterogeneous data sets [16, 32]. On the other hand, constraint-based methods involve a multiple testing problem [27]. The involved tests, whose results are inter-related in the process of constructing the causal graph, are usually performed independently, either accepting or rejecting the null hypothesis. Some of the testing results may conflict with each other. There exist some ways, e.g., by using logical encoding of independence constraints [17], to handle conflicts; However, how to determine the weights for different constraints remains an issue. Moreover, the power of statistical tests depends on the sample size, the number of conditioning variables, the variable dimensionality, etc., and it is usually hard to set the significance level of conditional independence tests in a principled way.

In contrast, score-based methods avoid some of the above issues. Instead of testing each (conditional) independence constraint independently with a binary decision, score-based methods evaluate the quality of candidate causal models with some score functions and output one or multiple graphs having the optimal score [14]. To calculate such scores, current prevailing methods assume a particular model class to describe causal mechanisms and data distributions, which narrows the scenarios where score-based methods are applicable. Widely-used score functions include the BIC/MDL score [25] and the BGe score [12] for linear-Gaussian models and the BDeu/BDe score for discrete data [5, 13]. More recently, some other score-based methods [2, 4, 18, 19, 26] and hybrid methods [9, 31] have been proposed; they exploit some assumptions of the model class. However, in real-world data, the assumed model class for causal mechanisms and data distributions may not hold. Such model misspecification may give misleading results. For example, in cases where underlying causal relations are highly nonlinear, the linear model assumption may lead to both spurious and missing edges, as we will illustrate in Section 2. If one discretizes continuous data and then apply the BDe or BDeu score, the discretization

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3220104

procedure may lose useful information in the data distribution and affect statistical efficiency in causal discovery.

To overcome the above limitations of score-based methods, it is desirable to develop new classes of score functions that apply to general causal mechanisms and data distributions. In this paper we introduce generalized score functions based on the characterization of general (conditional) independence relationships between random variables. Interestingly, this is achieved by defining suitable scores for a particular regression problem in Reproducing Kernel Hilbert Space (RKHS). Our framework provides a unified way to deal with a wide range of nonlinear causal relations and a wide class of data distributions, including non-Gaussian data and mixed continuous and discrete data. Moreover, we use kernel formulations in our approach and, as an advantage, it directly applies to variables with different dimensionalities.

Our contribution is mainly two-fold:

- We propose generalized score functions for causal discovery, which allow us to handle both linear and nonlinear causal relations and data with arbitrary distribuitons in an unified way and thus enable a wider range of applications of score-based causal discovery.
- We present the appealing properties of the proposed score functions. We show that by making use of the local score change in greedy equivalence search, it guarantees to find the Markov equivalence class which is consistent to the data generative distribution, even though our score is not equivalent (for different DAGs in the same equivalence class).

2 BACKGROUND AND MOTIVATION

The score-based method may exhibit some advantages over the constraint-based one [14]. However, for the score-based method, it is necessary to specify a proper model class. Current prevailing score-based methods usually make strong assumptions on the causal mechanism, as well as the data distribution.

In real-world data, we may not have precise information about the model class; causal relations can be linear or nonlinear, and data may have arbitrary distributions. If we misspecify the model class, it may result in both spurious edges and missing edges in the recovered causal graph. Figure 1 illustrates two cases when we use misspecified models.

In *Case* 1, variables X_1, X_2, and X_3 satisfy the following functional causal model: $X_1 = E_1, X_2 = 0.8(X_1 + X_1^2) + E_2$, and $X_3 = 0.8(X_2 + X_2^2) + E_3$, with $E_1, E_2, E_3 \sim \mathcal{N}(0, 0.5)$. If we use the BIC score under the linear-Gaussian model assumption, then with large enough samples, the graph corresponding to the optimal score will have a spurious edge between X_1 and X_3. Figure 1(a) shows the scatter plot of estimated noises $\hat{E}_1$ and $\hat{E}_3$, where $\hat{E}_1$ is the noise term of regressing X_1 on X_2, and $\hat{E}_3$ that of regressing X_3 on X_2. They are correlated because the influence of X_1 on X_3 can not be blocked by a *linear function* of X_2: although X_1 and X_3 are *nonlinearly* conditionally independent given X_2, the partial correlation between X_1 and X_3 given X_2 is nonzero, leading to the extra edge between X_1 and X_3.

In *Case* 2, we use X_1 to generate X_2 according to $X_1 = E_1$, and $X_2 = (\sin(X_1) + E_2)^2$, with $E_1, E_2 \sim \mathcal{N}(0, 0.5)$. If we use a linear-Gaussian model, we will miss the connection between X_1 and X_2.

Figure 1: (a) Scatter plot of the estimated noise $\hat{E}_1$ and $\hat{E}_3$; $\hat{E}_1$ and $\hat{E}_3$ are correlated. (b) Scatter plot of X_1 and X_2; they are uncorrelated.

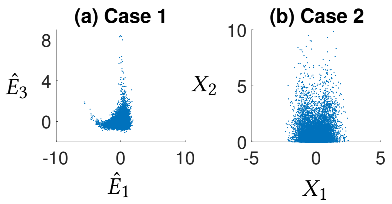

Figure 1(b) gives the scatter plot of X_1 and X_2; they are uncorrelated, although dependent.

Therefore, it is essential to develop score functions with wider applicability, that is, score functions that are able to handle general cases, without assuming particular model classes. Below we will show that this can be achieved by making use of the characterization of general conditional independence in the RKHS.

3 CHARACTERIZATION OF GENERAL CONDITIONAL INDEPENDENCE

For data generated by linear-Gaussian models, we can use partial correlation to capture conditional independence. Suppose we have three variables X, Y, and Z which satisfy $X \perp\!\!\!\perp Y|Z$. Let E_x be the error of regressing X on Z, and E_y the error of regressing Y on Z. Then we have the equivalence between $X \perp\!\!\!\perp Y|Z$ and $\mathrm{Cov}(E_x, E_y) = 0$. The latter implies that Y does not help in prediction of X, given Z already considered as a predictor. This also indicates that a properly defined score metric that is consistent in model selection, such as the BIC score, will use Z to predict X but not include Y as a predictor.

In the general case, how can we find such a score metric to verify conditional independence or dependence relations, without the prior information on causal mechanisms and data distributions? Interestingly, the general conditional independence can be characterized in the RKHS by cross-covariance operator.

Let us first give notations which will be used throughout the paper. We use X as a random variable, with domain $\mathcal{X}$. We define a RKHS $\mathcal{H}_\mathcal{X}$ on $\mathcal{X}$, with continuous feature mapping $\phi_\mathcal{X} : \mathcal{X} \to \mathcal{H}_\mathcal{X}$ and a measurable positive-definite kernel function $k_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which satisfies $k_\mathcal{X}(x, x') = \langle \phi_\mathcal{X}(x), \phi_\mathcal{X}(x') \rangle$. Let the probability law of X be denoted by P_X, and the space of square integrable functions with probability P_X by $L^2(P_X)$. We assume $\mathcal{H}_\mathcal{X} \subset L^2(P_X)$. Similar notations are applied to Y and Z.

Suppose that for random variable X, we have n observations $\mathbf{x} = (x^{(1)}, \cdots, x^{(n)})$. Let K_X represent the kernel matrix of sample $\mathbf{x}$, and the corresponding centralized kernel matrix is $\tilde{K}_X = HK_XH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathrm{T}$ with I and $\mathbf{1}$ being the $n \times n$ identity matrix and the vector of 1's, respectively. For a particular observation $x \in \mathcal{X}$, we represent its empirical feature map $\mathbf{k}_x$ as $\mathbf{k}_x = \left(k_\mathcal{X}(x^{(1)}, x), \cdots, k_\mathcal{X}(x^{(n)}, x)\right)^\mathrm{T}$.

Let $(\mathcal{H}_\mathcal{X}, k_\mathcal{X})$ and $(\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$ be the RKHSs over measurable spaces $\mathcal{X}$ and $\mathcal{Z}$, with measurable positive definite kernels $k_\mathcal{X}$ and $k_\mathcal{Z}$, respectively. For a random vector (X, Z) on $\mathcal{X} \times \mathcal{Z}$, the cross-covariance operator $\Sigma_{ZX}\colon \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Z}$ is defined by the relation

$$\langle f, \Sigma_{ZX} g \rangle_{\mathcal{H}_\mathcal{Z}} = E_{XZ}[g(X)f(Z)] - E_X[g(X)]E_Z[f(Z)]$$

for all $g \in \mathcal{H}_{\mathcal{X}}$ and $f \in \mathcal{H}_{\mathcal{Z}}$. If $Z = X$, Σ_{ZX} degenerates to the covariance operator Σ_{XX}, which is self-adjoint and positive definite. $\Sigma_{XX|Z}$ is a conditional covariance operator on $\mathcal{H}_{\mathcal{X}}$, defined as

$$\Sigma_{XX|Z} = \Sigma_{XX} - \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}.$$

Below we give the characterization of conditional independence in terms of conditional covariance operators in general cases.

Lemma 1 (Characterization of (conditional) independence with (conditional) covariance operators [10]). *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 $\mathcal{Z}$, respectively, with* **characteristic** *kernels. Let X, Y, and Z be random variables on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, respectively. Assume $E_{X|Z}[g(X)|Z = \cdot] \in \mathcal{H}_{\mathcal{Z}}$ and $E_{X|(Y,Z)}[g(X)|(Y,Z) = \cdot] \in \mathcal{H}_{\mathcal{Y}\mathcal{Z}}$, for all $g \in \mathcal{H}_{\mathcal{X}}$, where $\mathcal{H}_{\mathcal{Y}\mathcal{Z}}$ represents the direct product of $\mathcal{H}_{\mathcal{Y}}$ and $\mathcal{H}_{\mathcal{Z}}$. Then*

$$\Sigma_{XX|Z} - \Sigma_{XX|[Y,Z]} = 0 \Longleftrightarrow X \perp\!\!\!\perp Y|Z.$$

The kernel-based conditional independence tests have been proposed based on Lemma 1 [11, 35]. In this paper, we focus on score-based methods, and will show that some model selection criteria for regression in RKHS can capture general conditional independence, according to the above lemma.

Regression in RKHS. Suppose that we observe random variables X and Z over measurable spaces $\mathcal{X}$ and $\mathcal{Z}$, respectively. To encode general dependence relations between X and Z, we exploit a regression framework in the RKHS:

$$\phi_{\mathcal{X}}(X) = F(Z) + U \tag{1}$$

with $F : \mathcal{Z} \to \mathcal{H}_{\mathcal{X}}$, and U represents noise.

Let $\ddot{Z} := (Y, Z)$. Now let us consider the following two regression functions in RKHS.

$$\begin{aligned} \phi_{\mathcal{X}}(X) &= F_1(Z) + U_1, \\ \phi_{\mathcal{X}}(X) &= F_2(\ddot{Z}) + U_2. \end{aligned} \tag{2}$$

The regression in RKHS characterizes conditional independence relationships in the following way: as shown below, if $X \perp\!\!\!\perp Y|Z$, then we have

$$E_Z[\mathrm{Var}_{X|Z}[\phi_{\mathcal{X}}(X)|Z]] = E_{\ddot{Z}}[\mathrm{Var}_{X|Z}[\phi_{\mathcal{X}}(X)|\ddot{Z}]], \tag{3}$$

and vice versa. That is, it is not useful to incorporate Y as a predictor of X given Z, and thus, we prefer the former model in the equation set (2). Below we show that Eq. 3 can be derived from Lemma 1.

It has been shown that $\langle g, \Sigma_{XX|Z}g\rangle = E_Z[\mathrm{Var}_{X|Z}[g(X)|Z]]$ for all $g \in \mathcal{H}_{\mathcal{X}}$ [10]. From Lemma 1 we know that $\Sigma_{XX|Z} - \Sigma_{XX|\ddot{Z}} = 0 \Longleftrightarrow X \perp\!\!\!\perp Y|Z$. Thus, we have

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

We can write $\phi_{\mathcal{X}}(X)$ as $\phi_{\mathcal{X}}(X) = [\phi_1(X), \cdots, \phi_i(X), \cdots]^{\mathrm{T}}$, with $\mathrm{Cov}(\phi_i(X), \phi_j(X)) = 0$ for any $i \neq j$. Since $\phi_{\mathcal{X}}(X)$ is a feature map in $\mathcal{H}_{\mathcal{X}}$, for each component $\phi_i(X)$ of $\phi_{\mathcal{X}}(X)$, there exists a function $g \in \mathcal{H}_{\mathcal{X}}$ such that $g = \phi_i(X)$; i.e. Eq. 4 holds for any $\phi_i(X)$. Furthermore, based on the orthogonality between $\phi_i(X)$ and $\phi_j(X)$, we can derive Eq. 3.

Therefore, examining (conditional) independence relations in the general case can be seen as a model selection problem for appropriate regression tasks. Hence, causal structure learning can also be cast as such a model selection problem. In the next section, we develop score functions for the model selection, which are able to capture general conditional independence, without assuming any specific causal mechanisms and data distributions. Clearly, the score involves the likelihood and some measures of complexity.

4 GENERALIZED SCORE FUNCTIONS FOR CAUSAL DISCOVERY

It is well known that maximizing the likelihood function itself may lead to overfitting in structure learning. It is necessary to incorporate some complexity measures into the score function. In this section, we first define the likelihood function for regression in RKHS (Eq. 1), and based on it we then propose using cross-validated (CV) likelihood and marginal likelihood as score functions for structure learning.

4.1 Likelihood for Regression in RKHS

Suppose that we use a characteristic kernel such as the Gaussian kernel. In the formulation of the regression given in Eq. 1, the response variable, $\phi_{\mathcal{X}}(X)$, is in an infinite-dimensional space. As a consequence, we do not have a proper probability measure for $\phi_{\mathcal{X}}(X)$ and can not derive the likelihood function. Below we avoid this issue by considering a finite-dimensional projection of $\phi_{\mathcal{X}}(X)$ as the response variable.

Suppose that we have n observations $(\mathbf{x}, \mathbf{z}) = (x^{(1)}, z^{(1)}), \cdots, (x^{(n)}, z^{(n)})$ for the random vector (X, Z). For a particular observation $(x, z) \in (\mathcal{X}, \mathcal{Z})$, we map $\phi_{\mathcal{X}}(x)$ into its empirical feature map, which is an n-dimensional space:

$$\mathbf{k}_x = \begin{bmatrix} \langle k_{\mathcal{X}}(x^{(1)}, \cdot), \phi_{\mathcal{X}}(x)\rangle_{\mathcal{H}_{\mathcal{X}}} \\ \vdots \\ \langle k_{\mathcal{X}}(x^{(n)}, \cdot), \phi_{\mathcal{X}}(x)\rangle_{\mathcal{H}_{\mathcal{X}}} \end{bmatrix}.$$

With the property that functions in the RKHS are in the closure of linear combinations of the kernel at given points [24], i.e., $f(x) = \sum_{i=1}^{n} k(x^{(i)}, x)c_i$ with c_i being the weight, mapping $\phi_{\mathcal{X}}(x)$ into $\mathbf{k}_x$ does not cause loss of information. Thus, instead of using $\phi_{\mathcal{X}}(x)$ as the response variable in Eq. 1, we use $\mathbf{k}_x$. Then the regression in RKHS on finite observations is reformulated as

$$\mathbf{k}_x = \tilde{F}(z) + \tilde{U}. \tag{5}$$

We capture the regression error with the squared errors of $\tilde{U}$. That is, a Gaussian distribution is used for it. Now we can derive the log-likelihood function for the regression problem in Eq. 5. With the kernel trick, we represent the maximum log-likelihood with kernels, without explicitly resorting to the feature map. The maximum log-likelihood on finite data points is represented as

$$S_l(X, Z) = -\frac{n^2}{2}log(2\pi) - \frac{n}{2}\log\left|n\lambda^2\tilde{K}_X(\tilde{K}_Z + n\lambda I)^{-2}\tilde{K}_X\right| - \frac{n}{2}, \tag{6}$$

where λ is a regularization parameter. In this paper, we fix the kernel width, so we work on fixed feature spaces for all variables. See Appendix A1 for detailed derivations.

4.2 Generalized Score Functions

Based on the derived maximum log-likelihood for the regression in RKHS, we propose using CV log-likelihood and marginal log-likelihood as score functions for model selection, which is, in our scenario, causal structure learning. We assume that there is no feedback or hidden common cause in the underlying causal graph.

4.2.1 Cross-Validated Likelihood. Suppose that we have m variables, $X_1, \cdots, X_m$, which form a DAG $\mathcal{G}$, and n observations for the variables. We denote the current hypothetical DAG as $\mathcal{G}_h$. To do cross validation, we split the whole data set, denoted by D, into a training set and a test set and repeat this procedure Q times. The sample size of each training set is n_1, and that of corresponding test set is n_0, with $n_0 + n_1 = n$. Let $D_1^{(q)}$ and $D_0^{(q)}$ ($q = 1, \cdots, Q$) be the qth training set and qth test set, respectively. We further denote $D_{1,i}^{(q)}$ and $D_{0,i}^{(q)}$ for the corresponding data of variable X_i and its parents. One may use K-fold cross-validation or Monte-Carlo cross-validation.

Following the decomposable property,[1] the score of DAG $\mathcal{G}_h$ is represented as $S_{\mathrm{CV}}(\mathcal{G}_h; D) = \sum_{i=1}^{m} S_{\mathrm{CV}}(X_i, PA_i^{\mathcal{G}_h})$. For a particular variable X_i with parents $PA_i^{\mathcal{G}_h}$, its score is defined as

$$S_{\mathrm{CV}}(X_i, PA_i^{\mathcal{G}_h}) = \frac{1}{Q} \sum_{q=1}^{Q} \ell(\hat{\tilde{F}}_i^{(q)} | D_{0,i}^{(q)}), \tag{7}$$

where $S_{\mathrm{CV}}(X_i, PA_i^{\mathcal{G}_h})$ is the CV log-likelihood with X_i as the target variable and $PA_i^{\mathcal{G}_h}$ as predictors in the regression function in Eq. 5, $\hat{\tilde{F}}_i^{(q)}$ represents the regression function estimated from training data $D_{1,i}^{(q)}$, and $\ell(\hat{\tilde{F}}_i^{(q)} | D_{0,i}^{(q)})$ denotes the log-likelihood evaluated on the qth test set with the learned regression function $\hat{\tilde{F}}_i^{(q)}$.

Based on the regression function formulated in Eq. 5 and with kernel tricks, we represent $\ell(\hat{\tilde{F}}_i^{(q)} | D_{0,i}^{(q)})$ with kernel matrices :

$$\begin{aligned}
&\ell(\hat{\tilde{F}}_i^{(q)} | D_{0,i}^{(q)}) = \\
&-\tfrac{n_0^2}{2} log(2\pi) - \tfrac{n_0}{2} \log \left| n_1 \lambda^2 \tilde{K}_{X_i}^{1(q)} (\tilde{K}_{PA_i^{\mathcal{G}_h}}^{1(q)} + n_1 \lambda I)^{-2} \tilde{K}_{X_i}^{0(q)} \right| \\
&-\tfrac{1}{2} \mathrm{trace} \Big\{ \tfrac{1}{\lambda} \tilde{K}_{X_i}^{0(q)} \tilde{K}_{X_i}^{0(q)} + \tfrac{1}{\lambda} \tilde{K}_{PA_i^{\mathcal{G}_h}}^{0,1(q)} A_i^{\mathrm{T}} A_i \tilde{K}_{PA_i^{\mathcal{G}_h}}^{1,0(q)} \\
&-n_1 \tilde{K}_{PA_i^{\mathcal{G}_h}}^{0,1(q)} A_i^{\mathrm{T}} B_i A_i \tilde{K}_{PA_i^{\mathcal{G}_h}}^{1,0(q)} + 2 n_1 \tilde{K}_{X_i}^{0(q)} B_i A_i \tilde{K}_{PA_i^{\mathcal{G}_h}}^{1,0(q)} \\
&-\tfrac{2}{\lambda} \tilde{K}_{X_i}^{0(q)} A_i \tilde{K}_{PA_i^{\mathcal{G}_h}}^{1,0(q)} - n_1 \tilde{K}_{X_i}^{0(q)} B_i \tilde{K}_{X_i}^{0(q)} \Big\},
\end{aligned} \tag{8}$$

where $A_i = \tilde{K}_{X_i}^{1(q)} (\tilde{K}_{PA_i^{\mathcal{G}_h}}^{1(q)} + n_1 \lambda I)^{-1}$, $B_i = A_i \big(I + n_1 \lambda A_i^{\mathrm{T}} A_i \big)^{-1} A_i^{\mathrm{T}}$, λ is the regularization parameter, $\tilde{K}_{X_i}^{1(q)}$ denotes the centralized kernel matrix of the qth training set of $\mathbf{x}_i$, $\tilde{K}_{X_i}^{0(q)}$ denotes that of the qth test set of $\mathbf{x}_i$, and similar notations are used for other kernel matrices. See Appendix A2 for detailed derivations.

When using score functions for causal discovery, we care about whether the underlying causal graph or its equivalence class gives the optimal score. Specifically, here our concern is whether the score of a DAG model (1) increases as the result of adding any edge that eliminates an independence constraint that does not hold in the generative distribution, and (2) decreases as a result of adding any edge that does not eliminate such a constraint. It is about the property of *score local consistency.* More formally, we have the following definition of score local consistency [7].

Definition 1 (Score Local Consistency). *Let $\mathcal{G}$ be any DAG, and let $\mathcal{G}'$ be the DAG that results from adding the edge $X_i \rightarrow X_j$ on $\mathcal{G}$. Let D be the dataset from distribution $p(\cdot)$. A score function $S(\mathcal{G}; D)$ is locally consistent if the following two properties hold as the sample size $n \rightarrow \infty$:*

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

Here the graph which gives one more correct (conditional) independence constraint has a larger score.

For the regression problem one can define the effective dimension of the kernel space and the complexity of the regression function according to [6]. Then under mild conditions, the CV-likelihood score is locally consistent .

Lemma 2. *Suppose that the sample size of each test set n_0 satisfies*

$$n_0 \rightarrow \infty, \frac{n_0}{n} \rightarrow 0 \text{ as } n \rightarrow \infty,$$

and suppose that the regularization parameter λ satisfies

$$\lambda = O(n^{-\frac{b}{bc+1}}),$$

where b is a parameter of the effective dimension of the kernel space with $b > 1$, and c indicates the complexity of the regression function with $1 < c \leq 2$.

Then under conditions given in C1 & C2 (see Appendix A3), the CV likelihood under the regression framework in RKHS as a score function is locally consistent.

The detailed definition of b and c is shown in [6]. The condition that $n_0 \rightarrow \infty$ as $n \rightarrow \infty$ excludes leave-one-out cross validation. Although Lemma 2 requires that $\frac{n_0}{n} \rightarrow 0$ as $n \rightarrow \infty$, it has been shown that K-fold (e.g., K = 5 or 10) cross validation is a reasonable choice in practice [20]. The local consistency provides support for learning the causal structure which has the same independence constraints as the data generative distribution with the CV likelihood as the score. The proof is shown in Appendix A3.

It is known that there might be more than one DAG which share the same independence constraints. Now it comes to the question of whether those DAGs have the same score; if they have the same score, then the score function is said to be score equivalent [7].

Definition 2 (Score Equivalence). *Let D be the dataset from distribution $p(\cdot)$. A score function S is score equivalent if for any two DAGs $\mathcal{G}$ and $\mathcal{G}'$ which are in the same Markov equivalence class, we have $S(\mathcal{G}; D) = S(\mathcal{G}'; D)$.*

With the CV likelihood, we found that different DAGs in the same equivalence class may have different scores; i.e., it is not score equivalent. The reason is that the regression we use is nonlinear. It has been shown that with nonlinear relationships, models in the two directions, $X \rightarrow Y$ and $Y \rightarrow X$, may have different scores [15, 33, 34].

[1] A score function is *decomposable* if it can be written as a sum of measures, where each measure is a function of only one variable and its parents.

4.2.2 Marginal Likelihood. Alternatively, one may use marginal likelihood as a score function to avoid overfitting and infer the causal structure. The score function using the marginal likelihood over the hypothetical graph $\mathcal{G}_h$ is estimated by

$$S_M(\mathcal{G}_h; D) = \sum_{i=1}^{m} S_M(X_i, PA_i^{\mathcal{G}_h})$$

with $S_M(X_i, PA_i^{\mathcal{G}_h}) = \log p(X_i | PA_i^{\mathcal{G}_h}, \hat{\sigma}_i)$, where the hyperparameter $\hat{\sigma}_i^2$, as the noise variance, is learned by maximizing the marginal likelihood with gradient methods. Clearly, $S_M(\mathcal{G}_h; D)$ is decomposable. For a random variable X_i with parents $PA_i^{\mathcal{G}_h}$, the score function using the log marginal likelihood under the regression function in Eq. 5 can be written as

$$\begin{aligned} & S_M(X_i, PA_i^{\mathcal{G}_h}) \\ &= -\frac{1}{2}\mathrm{trace}\{\tilde{K}_{X_i}(\tilde{K}_{PA_i^{\mathcal{G}_h}} + \hat{\sigma}_i^2 I)^{-1}\tilde{K}_{X_i}\} \\ &- \frac{n}{2}\log|\tilde{K}_{PA_i^{\mathcal{G}_h}} + \hat{\sigma}_i^2 I| - \frac{n^2}{2}\log 2\pi. \end{aligned} \tag{9}$$

Here the kernel widths of variables are fixed in order to work on fixed RHKSs.

Similar to the CV likelihood, we investigate local consistency and score non-equivalence of the marginal likelihood. Lemma 3 shows that the marginal likelihood is locally consistent under mild conditions.

Lemma 3. *Under the condition that*

$$\lim_{n \to \infty} \frac{1}{6}(\sigma_i - \hat{\sigma}_i)^3 \frac{\partial^3 \log p(X_i | PA_i^{\mathcal{G}}, \hat{\sigma}_i)}{\partial \sigma_i^3} = 0, \tag{10}$$

and with a noninformative prior that $p(\sigma_i) = 1$ *over the neighborhood of* $\hat{\sigma}_i$*, the marginal likelihood under the regresson framework in RKHS as a score function is locally consistent.*

The condition (10) means that σ_i is close to $\hat{\sigma}_i$ as $n \to \infty$. Lemma 3 can be proved by making use of the Laplace method [8]. The proof is shown in Appendix A4.

When using the marginal likelihood as a score function, we found that, similar to the CV likelihood, different DAGs in the same Markov equivalence class may have different scores; i.e., the marginal likelihood under the regresson framework in RKHS is not score equivalent.

5 CAUSAL SEARCH PROCEDURE

Given a properly defined score function, we are now concerned with the search procedure that can give the optimal equivalence class asymptotically.

In Section 4.2 we have demonstrated that both CV likelihood and marginal likelihood are not score equivalent. From simulation results, we found that among DAGs within the underlying, true equivalence class, the DAG which has the same orientations as true causal directions usaually gives the highest score, for both CV and marginal likelihood. However, we do not have a theoretical justification for this observation yet. Therefore, we aim at searching for the optimal Markov equivalence class.

How can we search through the space of equivalence classes when the score is not equivalent? It may be the case that the DAG from the equivalence class with one less correct dependence or independence relation gives a higher score than a DAG from the other equivalence class does because of score non-equivalence. In the search procedure, if we compare the score of arbitray DAGs in two equivalence classes, in theory it is not guaranteed that we always introduce a correct dependence or independence relation.

Let us consider a simple example. Suppose that the ground truth is $X \to Y \leftarrow Z$. Further suppose that, during the search, the current equivalence class $\mathcal{E}_1$ is $X - Y \quad Z$, with the score S_1 estimated on the DAG $X \leftarrow Y \quad Z$. Now we attempt to move to the next equivalence class $\mathcal{E}_2$ $X \to Y \leftarrow Z$ by adding an edge between Y and Z and orienting the directions from X to Y and from Z to Y; denote the corresponding score by S_2. Note that since the score is not equivalent, and we do not have a guarantee on the consistency of the whole DAG, it is possible that S_1 is larger than S_2. How can we correctly go from $\mathcal{E}_1$ to $\mathcal{E}_2$, which is the underlying equivalence class?

A proper solution is to make use of the local score change in two adjacent equivalence classes, instead of comparing two arbitrary DAGs in them. Particularly, in the above example, we care about the local score change $S(Y; \{X, Z\}) - S(Y; X)$ from $\mathcal{E}_1$ to $\mathcal{E}_2$, which is positive, as guaranteed by local consistency. Fortunately, in the greedy equivalence search (GES) [7] procedure which is originally designed for equivalent score functions, it also searches for the maximum local score change, so we can prove that the GES search procedure with our scores is asymptotically optimal, even though the scores are not equivalent, as shown in the following propositions.

Proposition 1. *Assume that all conditions given in Lemma 2 hold. With the CV likelihood under the regression framework in RKHS as a score function and with the GES search procedure, it guarantees to find the Markov equivalence class which is consistent to the data generative distribution asymptotically.*

Proposition 2. *Assume that all conditions given in Lemma 3 hold. With the marginal likelihood under the regression framework in RKHS as a score function and with the GES search procedure, it guarantees to find the Markov equivalence class which is consistent to the data generative distribution asymptotically.*

Propositions 1 and 2 ensure that, with proper score functions and seach procedures, asymptotically the resulting Markov equivalence class has the same independence constraints as the data generative distribution. The proofs are shown in Appendix A5.

6 EXPERIMENTAL RESULTS

We applied the proposed generalized score functions, combined with GES search procedure, to both synthetic and real-world data sets to learn causal graphs, up to the Markov equivalence class.

6.1 Synthetic Data

Simulations To show the generality of the proposed generalized score functions, we generated different types of data, including

- continuous data: all variables in the causal graph are continuous;
- mixed continuous and discrete data: some variables are continuous, and some are discrete;

- multi-dimensional data: variables have different dimensions ranging from 1 to 5.

For each variable X_i, the data was generated according to the following functional causal model:

$$X_i = g_i(f_i(PA_i) + E_i),$$

where f_i represents the causal mechanism; it is randomly chosen from the *linear* function, *sin* function, *cos* function, *tanh* function, *logarithmic* function, and their combinations; g_i denotes post-nonlinear distortion in variable X_i, which is chosen from the *linear* function and the *exponential* function; E_i is the noise term, which is randomly chosen from *Gaussian*, *uniform*, and *gamma* distributions. Specifically, when f_i is the *logarithmic* function and g_i is the *exponential* function, it is a multiplicative-noise model.

We generated causal structures with different graph densities $dg = .2, .3, .4, .5, .6$, which are measured by the ratio of averaged degrees and the number of nodes. Each generated graph has 10 variables. In addition, we generated data with different sample sizes, $n = 500$ or 1000. For each setting (with a particular graph density, a particular data type, and a particular sample size), we generated 50 realizations; in total, there are $5 \times 3 \times 2$ settings.

We learned the causal structure by both of the proposed generalized score functions, the CV likelihood and the marginal likelihood under the regression framework in RKHS. We compared them with other score-based methods, including the kernel generalized variance (KGV) score [2] and the score using Spearman rank correlation to approximate mutual information proposed in [26], denoted by SC. Both are the state-of-the-art score-based methods to handle nonlinear causal relations and mixed continuous and discrete data. We also compared with the BIC score under the linear-Gaussian model assumption. We applied the GES search procedure, with the above score functions, to recover the causal graph up to the Markov equivalence class.

We also compared with constraint-based methods. Particularly, we used the kernel-based conditional independence (KCI) test to test for (conditional) independence relationships between variables [35]. The KCI test can handle nonlinear causal relations and data from arbitrary distributions. Since the search procedure may affect the results, we applied the widely-used PC search [28] and state-of-the-art search procedures, including the semi-interleaved HITON-MB search with symmetry correction [1] and the max-min Markov blanket (MM-MB) search with symmetry correction [1]. We did not compare with hybrid approaches, e.g., max-min hill-climbing (MMHC) [31], since it also uses constraint-based method for causal skeleton search.

For those methods that exploit kernels, we applied a Gaussian kernel and tried different kernel widths. We found that setting the kernel width to twice of median distance between points in input space gives the best results in most cases. For the CV likelihood, we tried 5-fold, 10-fold, and Monte-Carlo cross validation. We found that different types of cross validation give similar performance. Hence, in the following, we reported the results with the kernel width twice of median distance, 10-fold cross validation for the CV likelihood, and significance level 0.05 for independence tests in the constraint-based methods. The computational complexity for kernel-based method is $O(n^3)$. Specifically, our approach has the similar time complexity to others which expolit kernels, e.g.,MM-MB.

Figure 2 gives the F1 score [2] of the recovered causal skeleton in each setting with proposed generalized score functions (the CV likelihood and the marginal likelihood), compared with other score-based methods (BIC, KGV, and SC) and constriant-based methods (PC, HITON-MB, and MM-MB). We use the linear-Gaussian BIC score only for the continuous case, and the score-based SC is not applicable to the multi-dimenisonal case. The x-axis shows the graph density, measured by the ratio of averaged degrees to the number of nodes. The y-axis is the F1 score; higher F1 scores mean higher accuracies. Overall, we found that our proposed CV likelihood and marginal likelihood give the best accuracy in all settings, especially in the case of dense graphs, small sample sizes, and variables with multi-dimensionalities. The accuracy increases along with the sample size and decreases along with the graph density for all methods. More specifically, when the graph density increases, the accuracy of all score-based methods decreases much more slowly than the constraint-based methods. The constraint-based methods give better accuracy than the score-based KGV and SC when graphs are sparse, but they become worse when graphs get more dense, especially in the case of multi-dimensional variables. The reason may be that for constraint-based methods, the number of variables in the conditioning set increases along with the graph density, resulting in reduced power of conditional independence tests at a fixed significance level. In the mixed case and when sample size is large (Figure 2 (c.2)), the constraint-based methods are comparable to our proposed score functions. The performance of KGV and SC is not as good as that by others, probably for the following reasons: the KGV score does not consider interactions between multiple causes; the SC score uses the Spearman rank correlation between variables in the original space, and it relies on the assumption that the relationships between variables are monotonic.

We exploited another accuracy measurement, the normalized structural hamming distance (SHD) [31], to evaluate the difference between recovered Markov equivalence class and the true one, which considers recovered causal directions as well. The normalized SHD also demonstrates the efficacy of our metrics. Figure 3 gives the normalized SHD of the recovered MEC in each setting. The x-axis shows the graph density. The y-axis is the normalized SHD score; the lower the SHD score, the better accuracy. Overall, we found that the proposed CV likelihood and marginal likelihood give the best accuracy in all cases, especially in cases of dense graphs, small sample sizes, and variables with multi-dimensionalities, which is consistent with the results measured by F1 score.

The simulated testing results suggest that the proposed generalized score functions give best accuracy for almost all data types and all graph densities, especially in cases of dense graphs, small sample size, and variables with multi-dimensionalities.

Benchmark Datasets We also applied the proposed score functions to two benchmark discrete networks, i.e., CHILD network (20 variables) and SACHS network (11 variables), where all variables are discrete with cardinality ranging from 2 to 6. For each network, we randomly chose data points with sample size $n = 200, 500, 1000, 2000$ and repeated 50 times in each case. In addition

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

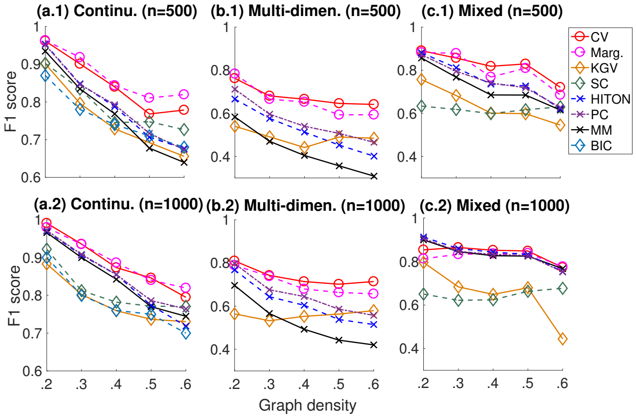

Figure 2: The F1 score of recovered causal graphs. (a.1) Continuous data with $n = 500$. (a.2) Continuous data with $n = 1000$. (b.1) Multi-dimensional data with $n = 500$. (b.2) Multi-dimensional data with $n = 1000$. (c.1) Mixed continuous and discrete data with $n = 500$. (c.2) Mixed continuous and discrete data with $n = 1000$. The x-axis is the graph density. The y-axis is the F1 score; higher F1 score means higher accuracy.

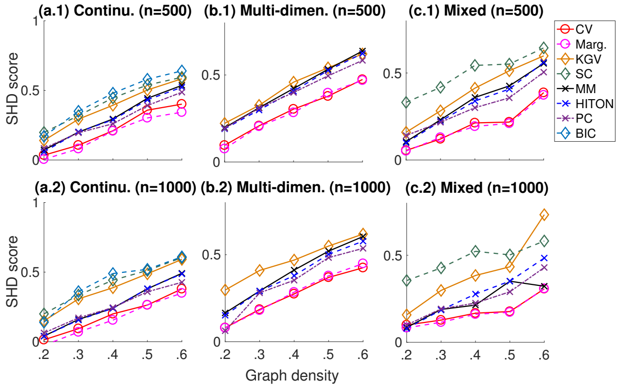

Figure 3: The normalized SHD of recovered causal graphs. The y-axis is the normalized SHD score; the lower SHD score means better accuracy.

to the methods used on the simulated data presented in the previous section, we compared our score functions with the well-known BDeu score which is designed for discrete data. For the BDeu score, the equivalent sample size is set to be $n' = 1$.

Figure 4 gives the F1 score of the recovered causal skeleton. The x-axis represents the sample size n. Overall, the accuracy increases as the sample size increases for all methods. The generalized score function with the marginal likelihood gives the best accuracy on SACHS at all sample sizes, while the BDeu score is slightly better than our proposed ones on CHILD. Both of the generalized score functions are better than all the constraint-based methods on both data sets and outperform BDeu on the network SACHS.

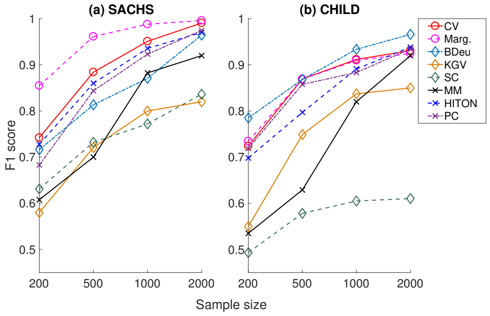

Figure 4: The F1 score of the recovered causal graphs on the two discrete networks. (a) CHILD network. (b) SACHS network.

6.2 Real-World Application

Archaeology data set. We then applied our methods to a real-world archaeology data set, collected by our collaborator Dr. Marlijn Noback. It contains eight variables with different dimensions, and the data are mixed continuous and discrete. The variables are: *Gender* (1 dimension, discrete), *Cranial size* (1 dimension, continuous), *Diet* (5 dimensions, discrete), *Paramasticatory behavior* (1 dimension, discrete), *Dental wear* (2 dimensions, mixed continuous and discrete), *Geographic location per population* (3 dimensions, discrete), *Climate per population* (6 dimensions, discrete), and *Cranial shape differentiation* (4 dimensions, continuous). The sample size n is 255.

Given that in simulation studies our proposed score functions have the best causal discovery performance, we applied them to the archaeology data set to identify causal relationships between the archaeology-related variables. The settings for kernels and cross validation are the same as those applied to synthetic data.

Figure 5 shows the recovered causal graph by the CV likelihood and the marginal likelihood. The solid lines are shared edges from both of them. The dashed edges are recovered only by the CV likelihood, and the dotted edges are recovered only by the marginal likelihood. The two resulted graphs mainly differ in the causal edges out from *Geographical location* and *Climate*. We consider the union of graphs recovered from these two score functions. We found that *Geographical location* and *Climate* are main causes of other variables. Both of them influence *Diet*, *Paramasticatory behavior*, *Cranial size*, and *Cranial shape differentiation*. *Gender* directly influences *Cranial size*. *Paramasticatory behavior* influences *Cranial shape differentiation*. The recovered causal relations are in accordance with our common understandings and domain knowledge. For example, *Climate* → *Cranial size* matches with Bergmann's rule that body size is large in cold climates and small in warm climates [21]. *Geography location* → *Shape differentiation* reflects genetic processes of isolation by distance [3]. *Gender* → *Cranial size* may reflect the fact that the male generally have bigger heads than the female [23]. *Climate* → *Diet* coincides with the phenomenon that humidity and temperature influence the type of plants and animals that are around to eat and influence whether one can do agriculture [30]. The results illustrate the effectiveness of our proposed methods in inferring causal relations from real-world, complex data.

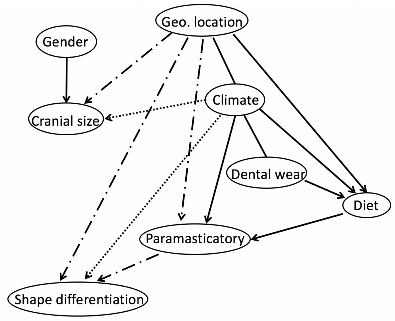

Figure 5: Recovered causal graph from Archaeology data set. The solid lines are shared edges from the CV likelihood and the marginal likelihood. The dashed edges are recovered only by CV likelihood, and the dotted edges are recovered only by marginal likelihood.

7 CONCLUSION

In this paper, we proposed generalized score functions for causal discovery, which can handle nonlinear causal relations and data with arbitrary distributions and dimensionalities in a unified way. We showed that they are score local consistent under mild conditions. With the GES search procedure, it guarantees to find the underlying Markov equivalence class asymptotically although the score equivalence is not satisfied. A line of our future research is to improve the computational efficiency of our approach and extend it to cases where there are feedbacks and confounders in the underlying causal graph.

ACKNOWLEDGEMENTS

We would like to acknowledge the support from United States Air Force under Contract No. FA8650-17-C-7715 and National Institutes of Health under Contract No. NIH-1R01EB022858- 01, FAIN-R01EB022858, NIH-1R01LM012087, NIH-5U54HG008540-02, and FAIN-U54HG008540. The content is solely the responsibility of the authors and does not necessarily represent the official views of the United States Air Force or the National Institutes of Health. We appreciate the comments from anonymous reviewers, which helped to improve the paper.

APPENDIX

A1: DERIVATION OF LIKELIHOOD

Let X be the target variable and Z the set of regressors. Suppose that for the random vecrtor (X, Z) there are n observations, with $(\mathbf{x}, \mathbf{z}) = (x^{(1)}, z^{(1)}), \cdots, (x^{(n)}, z^{(n)})$. For a particular observation (x, z), the formulation of regression in RKHS on finite sample size can be written as

$$\mathbf{k}_x = \tilde{F}(z) + \tilde{U}.$$

Then with kernel ridge regression on n observations, we have

$$\hat{\tilde{F}}_Z = \tilde{K}_X(\tilde{K}_Z + n\lambda I)^{-1}K_Z,$$

where λ is the regularization parameter. Thus the estimated covariance matrix of the residual is $\hat{\Sigma} = \frac{1}{n}(\tilde{K}_X - \hat{\tilde{F}}_Z)(\tilde{K}_X - \hat{\tilde{F}}_Z)^{\mathrm{T}} = n\lambda^2\tilde{K}_X(\tilde{K}_Z + n\lambda I)^{-2}\tilde{K}_X$.

Therefore, the maximal value of log-likelihood is represented as

$$\begin{aligned}
&S_l(X, Z) \\
&= -\frac{n^2}{2}log(2\pi) - \frac{n}{2}\log|\hat{\Sigma}| - \frac{1}{2}\mathrm{trace}\{(\tilde{K}_X - \hat{\tilde{W}}\Phi_Z)^T\hat{\Sigma}^{-1}(\tilde{K}_X - \hat{\tilde{W}}\Phi_Z)\} \\
&= -\frac{n^2}{2}log(2\pi) - \frac{n}{2}\log\left|n\lambda^2\tilde{K}_X(\tilde{K}_Z + n\lambda I)^{-2}\tilde{K}_X\right| - \frac{n}{2}.
\end{aligned}$$

In practice the inverse of $(\tilde{K}_Z + n\lambda I)$ can be calculated efficiently by Cholesky decomposition with $(\tilde{K}_Z + n\lambda I) = LL^{\mathrm{T}}$, where L is a lower triangular matrix, and thus, $(\tilde{K}_Z + n\lambda I)^{-1} = L^{-\mathrm{T}}L^{-1}$.

A2: DERIVATION OF CROSS-VALIDATED LIKELIHOOD

Let X be the target variable and Z be the set of regressors. We give the derivation of the cross-validated likelihood $S_{\mathrm{CV}}(X, Z)$, where $S_{\mathrm{CV}}(X, Z) = \frac{1}{Q}\sum_{q=1}^{Q}\ell(\hat{\tilde{F}}^{(q)}|D_{0,i}^{(q)})$. We consider two cases: Z is not empty and Z is empty.

Z **is not empty.** We first learn $\tilde{F}$ on the qth training set with kernel ridge regression, with

$$\hat{\tilde{F}}^{(q)} = \tilde{K}_X^{1(q)}(\tilde{K}_Z^{1(q)} + n_1\lambda I)^{-1}\tilde{K}_Z^{1(q)}.$$

Then the likelihood evaluated on the qth test set with the learned regression function $\hat{\tilde{F}}^{(q)}$ is derived as follows:

$$\begin{aligned}
&\ell(\hat{\tilde{F}}^{(q)}|D_{0,i}^{(q)}) \\
&= -\frac{n_0^2}{2}log(2\pi) - \frac{n_0}{2}\log|\hat{\Sigma}^{(q)}| - \frac{1}{2}\mathrm{tr}\Big\{(\tilde{K}_X^{0(q)} - \hat{\tilde{F}}^{(q)}(\hat{\Sigma}^{(q)})^{-1}(\tilde{K}_X^{0(q)} - \hat{\tilde{K}}^{(q)})\Big\} \\
&\doteq -\frac{n_0^2}{2}log(2\pi) - \frac{n_0}{2}\log|\hat{\Sigma}^{(q)}| - \frac{1}{2}\mathrm{trace}\Big\{(\tilde{K}_X^{0(q)} - \hat{\tilde{F}}^{(q)})^T(\hat{\Sigma}^{(q)} + \lambda I)^{-1} \\
&\qquad\qquad (\tilde{K}_X^{0(q)} - \hat{\tilde{F}}^{(q)})\Big\} \\
&= -\frac{n_0^2}{2}log(2\pi) - \frac{n_0}{2}\log\left|n_1\lambda^2\tilde{K}_X^{1(q)}(\tilde{K}_Z^{1(q)} + n_1\lambda I)^{-2}\tilde{K}_X^{1(q)}\right| \\
&\quad - \frac{1}{2}\mathrm{trace}\Bigg\{\frac{1}{\lambda}\tilde{K}_X^{0(q)}K_X^{0(q)} + \frac{1}{\lambda}\tilde{K}_Z^{0,1(q)}A^{\mathrm{T}}A\tilde{K}_Z^{1,0(q)} - \frac{2}{\lambda}\tilde{K}_X^{0(q)}AK_Z^{1,0(q)} \\
&\quad - n_1\tilde{K}_X^{0(q)}B\tilde{K}_X^{0(q)} - n_1\tilde{K}_Z^{0,1(q)}A^TBA\tilde{K}_Z^{1,0(q)} + 2n_1\tilde{K}_X^{0(q)}BA\tilde{K}_Z^{1,0(q)}\Bigg\},
\end{aligned}$$

where $A = \tilde{K}_X^{1(q)}(\tilde{K}_Z^{1(q)} + n_1\lambda I)^{-1}$, $B = A\Big(I + n_1\lambda A^{\mathrm{T}}A\Big)^{-1}A^{\mathrm{T}}$. The third equality uses the Woodbury identity.

Z **is empty.** Now we deal with the case when Z is empty. The likelihood evaluated on the qth test set with the learned regression function $\hat{\tilde{F}}^{(q)}$ is derived as follows:

$$\begin{aligned}
&\ell(\hat{\tilde{F}}^{(q)}|D_{0,i}^{(q)})\\
&= -\frac{n_0^2}{2} log(2\pi) - \frac{n_0}{2}\log|\hat{\Sigma}^{(q)}| - \frac{1}{2}\text{trace}\Big\{\tilde{K}_X^{1(q)}(\hat{\Sigma}^{(q)})^{-1}\tilde{K}_X^{0(q)}\Big\}\\
&\doteq -\frac{n_0^2}{2} log(2\pi) - \frac{n_0}{2}\log|\hat{\Sigma}^{(q)}| - \frac{1}{2}\text{trace}\Big\{\tilde{K}_X^{0(q)}(\hat{\Sigma}^{(q)} + \lambda I)^{-1}\tilde{K}_X^{0(q)}\Big\}\\
&= -\frac{n_0^2}{2}\log(2\pi) - \frac{n_0}{2}\log|\frac{1}{n_1}\tilde{K}_X^{1(q)}\tilde{K}_X^{1(q)}| - \frac{1}{2}\text{trace}\Bigg\{\frac{1}{\lambda}\tilde{K}_X^{0(q)}\tilde{K}_X^{0(q)}\\
&- \frac{1}{n_1\lambda^2}\tilde{K}_X^{0(q)}\tilde{K}_X^{1(q)}(I + \frac{1}{n_1\lambda}\tilde{K}_X^{1(q)}\tilde{K}_X^{1(q)})^{-1}\tilde{K}_X^{1(q)}\tilde{K}_X^{0(q)}\Bigg\}.
\end{aligned}$$

A3: PROOF OF LEMMA 2

We define

$$S_{\bar{o}}(X_i, PA_i^{\mathcal{G}}) := n_0 \cdot \frac{1}{Q}\sum_{q=1}^{Q}\ell(\hat{\tilde{F}}_i^{(q)}|P_i)$$

and

$$S_o(X_i, PA_i^{\mathcal{G}}) := n_0 \cdot \ell(\hat{\tilde{F}}_i|P_i)$$

to represent the optimal benchmark, where P_i denotes the true distribution of X_i, $\ell(\hat{\tilde{F}}_i^{(q)}|P_i)$ is evaluated on the true distributiom with the model being fit to the qth training set, while in $\ell(\hat{\tilde{F}}_i|P_i)$ the model is being fit to the entire dataset with sample size n. Furthermore, we introduce

$$S_*(X_i) := \int \log(f_i)\,\mathrm{d}P_i,$$

where f_i is the true density function of X_i corresponding to P_i.

Suppose that given data D, we have a set of candidate models, denoted by $\mathcal{M} = \{\mathcal{G}^{(1)}, \cdots, \mathcal{G}^{(k)}\}$ with cardinality k. Define

$$\hat{\kappa} = arg \max_{\kappa=1,\cdots,k} S_{\text{CV}}(X_i, PA_i^{\mathcal{G}^{(\kappa)}}),$$

$$\bar{\kappa} = arg \max_{\kappa=1,\cdots,k} S_{\bar{o}}(X_i, PA_i^{\mathcal{G}^{(\kappa)}}),$$

and

$$\mathring{\kappa} = arg \max_{\kappa=1,\cdots,k} S_o(X_i, PA_i^{\mathcal{G}^{(\kappa)}});$$

i.e., $\hat{\kappa}$ is the model selected by S_{CV}, $\bar{\kappa}$ is the model selected by $S_{\bar{o}}$, and $\mathring{\kappa}$ is the one selected by the benchmark S_o.

We give two mild conditions, which are used in Lemma 2.

Mild Conditions:

C1. There exist $\epsilon > 0$ and $C < \infty$, so that the likelihood $L(\hat{\tilde{F}}|D_i) \in (\epsilon, C)$ almost surely for all $\mathcal{G}^{(\kappa)} \in \mathcal{M}$ $(\kappa = 1, \cdots, k)$.

C2. The relation between $S_o(X_i, PA_i^{\mathcal{G}^{(\mathring{\kappa})}})$ and $S_{\bar{o}}(X_i, PA_i^{\mathcal{G}^{(\bar{\kappa})}})$ satisfies

$$\frac{S_o(X_i, PA_i^{\mathcal{G}^{(\mathring{\kappa})}}) - S_*(X_i)}{S_{\bar{o}}(X_i, PA_i^{\mathcal{G}^{(\bar{\kappa})}}) - S_*(X_i)} \xrightarrow{p} 1, \text{ for } n \to \infty.$$

Proof. Under condition C1, it has been shown [20] that if

$$\frac{\log(k)}{n_0 \cdot \Big(S_{\bar{o}}(X_i, PA_i^{\mathcal{G}^{(\bar{\kappa})}}) - S_*(X_i)\Big)} \xrightarrow{p} 0, \text{ for } n \to \infty, \tag{A1}$$

then

$$\frac{S_{\bar{o}}(X_i, PA_i^{\mathcal{G}^{(\hat{\kappa})}}) - S_*(X_i)}{S_{\bar{o}}(X_i, PA_i^{\mathcal{G}^{(\bar{\kappa})}}) - S_*(X_i)} \xrightarrow{p} 1. \tag{A2}$$

For Eq. A1 to hold, it requires $n_0 \to \infty$ as $n \to \infty$, excluding the case of leave-one-out cross validation. Eq. A2 says that the model selected by S_{CV} has the same performance as that selected by $S_{\bar{o}}$ in probability.

Furthermore, under the condition $\frac{n_0}{n} \to 0$ as $n \to \infty$ and condition C2, the model $\hat{\kappa}$ selected by S_{CV} satisfies the following property [20] :

$$\frac{S_{\bar{o}}(X_i, PA_i^{\mathcal{G}^{(\hat{\kappa})}}) - S_*(X_i)}{S_o(X_i, PA_i^{\mathcal{G}^{(\mathring{\kappa})}}) - S_*(X_i)} \xrightarrow{p} 1, \text{ for } n \to \infty. \tag{A3}$$

Eq. A3 says that the model $\hat{\kappa}$ selected by S_{CV} performs as well as the benchmark selector S_o on the whole sample size, as $n \to \infty$.

Furthermore, [6] has shown that as the regularization parameter λ in the kernel ridge regression satisfies $\lambda = O(n^{-\frac{b}{bc+1}})$, the estimation of $\tilde{F}$ is optimal in a minmax sense, so is S_o.

Therefore, S_{CV} is locally consistent; i.e., it chooses the correct model with probability 1.

□

A4: PROOF OF LEMMA 3

Proof. Since we work on fixed feature spaces, σ_i is the only (hyperparameter) parameter when using the marginal likelihood $S_M(X_i, PA_i^{\mathcal{G}_h})$ as a score function. By the Laplace method, we can derive

$$\log p(X_i|PA_i^{\mathcal{G}_h}) \approx \log L_i(\hat{\sigma}_i|D) - \frac{1}{2} \cdot log(\frac{n}{2\pi}). \tag{A4}$$

From Eq. A4, we can see that $\log p(X_i|PA_i^{\mathcal{G}_h})$ is written as the form of Bayesian information criterion (BIC). Since BIC is consistent, $\log p(X_i|PA_i^{\mathcal{G}_h})$ as a score function is consistent. Furthermore, since for a fixed dataset D with sample size n, the second term in Eq. A4 is a constant, we can directly use $\log p(X_i|PA_i^{\mathcal{G}_h}, \hat{\sigma}_i)$ as a score function. Therefore, S_M is locally consistent.

□

A5: PROOF OF PROPOSITION 4 AND PROPOSITION 5

We first consider the proof of Lemma 4. It consists of two parts: the proof of the forward phase and that of the backward phase of GES. In the forward phase, the resulting equivalence class $\mathcal{E}_f$ contains underlying distribution p; i.e., all independence constraints holding in $\mathcal{E}_f$ hold in p. It has been proved by making use of local consistency of score functions in [7].

We focus on showing that the backward phase is guaranteed to find a perfect map of p even when the score is not equivalent and the number of parameters in the same equivalence class is not the same.

Proof. Let $\mathcal{E}_b$ denote the equivalence class resulting from the backward phase of GES, and let $\mathcal{E}^*$ be the perfect map of p; i.e., all independence constraints in $\mathcal{E}^*$ are in p, and vice versa. Now we show that as the sample size $n \to \infty$, $\mathcal{E}_b = \mathcal{E}^*$.

First we show that the equivalence class $\mathcal{E}$ results from each step in the backward phase contains p. Consider a move from $\mathcal{E}$ to $\mathcal{E}^{-}(\mathcal{E})$ by applying Delete($X_i, X_j, \mathbf{H}$) (see the definition in [7]), where $\mathcal{E}$ contains p and $\mathcal{E}^{-}(\mathcal{E})$ does not contain p. Let $\mathcal{G} \in \mathcal{E}$ and $\mathcal{G}' \in \mathcal{E}^{-}(\mathcal{E})$ with the difference in $X_i \to X_j$. From the fact that the score functions are locally consistent, the local score change $\Delta S < 0$, so $S(\mathcal{G}; D) > S(\mathcal{G}'; D)$. The attempted move from $\mathcal{E}$ to $\mathcal{E}^{-}(\mathcal{E})$ will be rejected.

Next we show that the backward phase will not terminate with some suboptimal equivalence class $\mathcal{E}$; that is, there are no independence constraints which containing in p are not in $\mathcal{E}$.

Suppose that the backward phase terminates with some suboptimal equivalence class $\mathcal{E}$, and there is one more edge $X_i \to X_j$ or $X_i - X_j$ in $\mathcal{E}$ than in $\mathcal{E}^*$. According to local consistency, and the calculation of local score change with Delete operator, ΔS from $\mathcal{E}$ to $\mathcal{E}^*$ is positive; that is, the score of $\mathcal{E}^*$ is larger than that of $\mathcal{E}$. Hence it will move to $\mathcal{E}^*$. It contradicts with the assumption that the backward phase terminates with some suboptimal equivalence class. Therefore, the resulting equivalence class in the backward phase is a perfect map of p.

The proof does not require score equivalence. □

The proof of Proposition 5 is the same as that of Proposition 4, since the marginal likelihood also satisfies local consistency.

REFERENCES

[1] C. F. Aliferis, A. R. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. 2010. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research* 11 (2010), 171–234.

[2] F. R. Bach and M. I. Jordan. 2002. Learning graphical models with Mercer kernels. *Advances in Neural Information Processing Systems* (2002), 1009–1016.

[3] T. E. Bakken, A. M. Dale, and N. J. Schork. 2011. A Geographic Cline of Skull and Brain Morphology among Individuals of European Ancestry. *Hum Hered* 72(1) (2011), 35–44.

[4] P Bühlmann, J. Peters, and J. Ernest. 2014. CAM: Causal Additive Models, high-dimensional order search and penalized regression. *Annals of Statistics* 42(6) (2014), 2526–2556.

[5] W. Buntine. 1991. Theory refinment on Bayesian networks. *Uncertainty in Artificial Intelligence* (1991), 52–60.

[6] A. Caponnetto and E. De Vito. 2006. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics* (2006).

[7] D. M. Chickering. 2003. Optimal Structure Identification With Greedy Search. *Journal of Machine Learning Research* 3 (2003), 507–554.

[8] D. M. Chickering and D. Heckerman. 1997. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine Learning* 29 (1997), 181–212.

[9] T. Claassen and T. Heskes. 2012. A Bayesian approach to constraint based causal inference. *Uncertainty in Artificial Intelligence* (2012), 207–216.

[10] K. Fukumizu, F. R. Bach, and M. I. Jordan. 2004. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research* 5 (2004), 73–79.

[11] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. 2007. Kernel measures of conditional dependence. *NIPS* 11 (2007), 489–496.

[12] D. Geiger and D. Heckerman. 1994. Learning Gaussian networks. *In Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence* (1994), 235 –243.

[13] D. Heckerman, D. Geiger, and D.M. Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20 (1995), 197–243.

[14] D. Heckerman, C. Meek, and G. Cooper. 2006. A Bayesian approach to causal discovery. *Innovations in Machine Learning* (2006), 1–28.

[15] P. Hoyer, D. Janzing, J. Mooji, Peters J., and B. Schölkopf. 2009. Nonlinear causal discovery with additive noise models. *NIPS* (2009).

[16] B. Huang, K. Zhang, J. Zhang, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. 2017. Behind Distribution Shift: Mining Driving Forces of Changes and Causal Arrows. *ICDM* (2017), 913–918.

[17] A. Hyttinen, F. Eberhardt, and M. Järvisalo. 2014. Constraint-based causal discovery: Conflict resolution with answer set programming. *Uncertainty in Artificial Intelligence* (2014), 340–349.

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

[19] S. Imoto, T. Goto, and S. Miyano. 2002. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pacific Symposium on Biocomputing* (2002), 175–186.

[20] M. V. D. Laan, S. Dudoit, and S. Keles. 2004. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology* 3(1) (2004), 1–23.

[21] S. Meiri and T. Dayan. 2003. On the validity of Bergmann's rule. *Journal of Biogeography* 30(3) (2003), 331–351.

[22] J. Pearl. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press New York.

[23] A. N.V. Ruigrok, G. S. Khorshidi, M. Lai, S. B. Cohen, M. V. Lombardo, R. J. Tait, and J. Suckling. 2014. A meta-analysis of sex differences in human brain structure. *Neuroscience and Biobehavioral Reviews* 39 (2014), 34–50.

[24] B. Schölkopf and A. J. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA.

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

[26] E. Sokolova, P. Groot, T. Claassen, and T. Heskes. 2014. Causal discovery from databases with discrete and continuous variables. *Workshop on Probabilistic Graphical Models* (2014), 442–457.

[27] P. Spirtes. 2010. Introduction to Causal Inference. *Journal of Machine Learning Research* 11 (2010), 1643–1662.

[28] P. Spirtes, C. Glymour, and R. Scheines. 1993. *Causation, Prediction, and Search.* Spring-Verlag Lectures in Statistics.

[29] P. Spirtes and K. Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics* 3(3) (2016).

[30] M. Springmann, D. Mason-DCroz, S. Robinson, P. Ballon, T. Garnett, and C. Godfray. 2016. The global and regional health impacts of future food production under climate change. *The Lancet* 387 (10031) (2016), 1937–1946.

[31] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning* 65(1) (2006), 31–78.

[32] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. *IJCAI* (2017).

[33] K Zhang and A. Hyvärinen. 2009. Causality discovery with additive disturbances: An information-theoretical perspective. *Machine learning and knowledge discovery in databases* (2009), 570–585.

[34] K. Zhang and A. Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. *UAI* (2009), 647–655.

[35] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. 2011. Kernel-based conditional independence test and application in causal discovery. *Uncertainty in Artificial Intelligence* (2011), 804–813.

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