Causal Discovery by Kernel Intrinsic Invariance Measure

Zhitang Chen  Shengyu Zhu  Yue Liu  Tim Tse
Huawei Noah’s Ark Lab
{chenzhitang2,zhushengyu,liuyue52,tim.tse}@huawei.com

Abstract

Reasoning based on causality, instead of association has been considered as a key ingredient towards real machine intelligence. However, it is a challenging task to infer causal relationship/structure among variables. In recent years, an Independent Mechanism (IM) principle was proposed, stating that the mechanism generating the cause and the one mapping the cause to the effect are independent. As the conjecture, it is argued that in the causal direction, the conditional distributions instantiated at different value of the conditioning variable have less variation than the anti-causal direction. Existing state-of-the-arts simply compare the variance of the RKHS mean embedding norms of these conditional distributions. In this paper, we prove that this norm-based approach sacrifices important information of the original conditional distributions. We propose a Kernel Intrinsic Invariance Measure (KIIM) to capture higher order statistics corresponding to the shapes of the density functions. We show our algorithm can be reduced to an eigendecomposition task on a kernel matrix measuring intrinsic deviance/invariance. Causal directions can then be inferred by comparing the KIIM scores of two hypothetic directions. Experiments on synthetic and real data are conducted to show the advantages of our methods over existing solutions.

1 Introduction

Recent breakthrough in deep learning has been significantly advancing Artificial Intelligence (AI). We witness great success of deep learning in many applications such as image classification, image recognition, speech recognition, natural language processing etc. Deep learning methods, or specifically deep neural networks have become the dominant approach for machine learning and AI and thus attracts tremendous amount of attention from both the academia and the industry. However, there are still a number of open challenges remained to tackle for deep learning. To name a few, heavy demand on labeled data, bad generalizability, vulnerable against adversarial attacks and lack of interpretability of deep learning methods are the most notorious ones. Recently, it is advocated in the AI community that causality might be one of the tools to solve the aforementioned open problems. It has been argued that causality, instead of “superficial association” is invariant cross domain. Machine learning algorithms that learn, and utilize the causal relationship amongst variables provide better generalization performance, robustness against adversarial attacks and better interpretability. Besides the area of machine learning and AI [1, 16, 4, 10], causal discovery also play an important role in economics, sociology, bioinformatic and medical science etc.

However, how to unveil the causal relationship among variables from pure observational (or post-intervention) data is challenging. A bunch of methods have been proposed in the past three decades including Bayesian network [13], Structural Equation Models (SEM) [13, 7, 5]. However, these methods have their limitations. For example, Bayesian networks via constained-based approach or score-based approach are not able to fully identify the ground-truth graphs but only up to “Markov

Preprint. Under review.
where the IM principle is that nature is parsimonious in the sense that the mechanism generating the cause and the mechanism mapping the cause to the effects are independent, i.e. the probability distribution of the cause \( P(X) \) and the conditional distribution mapping the cause to the effect \( P(Y|X) \) contain no information of each other. It has been shown that the factorization of the joint distribution according to the causal direction usually yield simpler terms than that in the anti-causal direction \([22, 9]\), i.e.

\[
K(P(X)) + K(P(Y|X)) \leq K(P(Y)) + K(P(X|Y)),
\]

where \( K(\cdot) \) denotes the Kolmogorov complexity which is essentially not computable. Researchers have been proposing computable metrics including RKHS norm \([22, 5]\), Minimal Description Length \([2]\) etc., to mimic the Kolmogorov complexity in order to derive a practical algorithm for pairwise causal discovery. Our method proposed in this paper falls into this category. According to the IM principle, the conditional distribution \( P(Y|X) \) does not depend on \( P(X) \) which naturally leads to the conjecture that intrinsic information, e.g. higher order central moments that characterize the shape of \( P(Y|X = x) \) does not essentially depend on the value of \( x \). In this paper, we prove that existing state-of-the-art norm-based approach along this direction is not sufficient as it sacrifices important information of the original conditional distributions. Instead, we propose a Kernel Intrinsic Invariance Measure (KIIM) to capture the intrinsic invariance of the conditional distribution, i.e. the higher order statistics corresponding to the shapes of the density functions. We show our algorithm can be reduced to an eigen-decomposition task on a kernel matrix measuring intrinsic deviance/invariance.

The rest of the paper is organized as follows: in Sec\([2]\) we introduce the basic idea of a recent state-of-the-art method named Kernel Deviance Measure and its limitation; in Sec\([3]\) we give a brief introduction to Reproducing Kernel Hilbert Space (RKHS) embeddings which serve as the tool of our method; in Sec\([4]\) we give a rigorous justification of the limitation of existing methods and then show how our proposed method address those issues; in Sec\([5]\) we verify the effectiveness of our proposed method followed by a conclusion in Sec\([6]\).

## 2 Related Work

Recently, authors in \([11]\) proposed an idea which exploits the variation of the conditional distribution of the hypothetic effect given the hypothetic cause. They argued that there is less variability in the causal direction than that in the anti-causal direction. An motivating example that is used in \([11]\) as follows. Suppose we have two random variables that follow the generating mechanism as \( y = x^3 + x + \epsilon \), where \( \epsilon \sim \mathcal{N}(0, 1) \). As illustrated in Fig.\([1]\), it is obvious that the conditional distribution \( p(Y|x) \) instantiated at different value \( x \) are almost identical except for the location; however in the anti-causal direction, the conditional distribution \( p(X|y) \) instantiated at different value \( y \) have significant structural variation including the number of modes, skewness, kurtosis etc. This piece of structural variation in conditional distributions leads to the so-called “cause-effect asymmetry” for causal discovery. The basic idea is to investigate how invariant the conditional distribution (instantiated at different values) is and one prefers the direction with less variation or in other words, more invariance. To achieve it, they proposed the following Kernel Deviance Measure:

\[
S_{x \rightarrow y} := 1/n \sum_i (||\mu_{Y|x_i}||_{\mathcal{H}_Y} - 1/n \sum_j ||\mu_{Y|x_i}||_{\mathcal{H}_Y})^2,
\]

where \( \mathcal{H} \) is a Reproducing Kernel Hilbert Space (RKHS) entailed by a positive definite kernel \( k(\cdot, \cdot) \) and \( \mu_{Y|x_i} \) is the kernel mean embedding of the conditional distribution \( p(y|x) \) instantiated at \( x_i \).
The causal discovery rule is straightforward by comparing the scores, i.e. \( x \rightarrow y \), if \( S_{x \rightarrow y} < S_{y \rightarrow x} \), \( y \rightarrow x \), if \( S_{x \rightarrow y} > S_{y \rightarrow x} \), otherwise no conclusion is drawn.

Positive results were reported in [11] which suggests that causal discovery via invariance is a promising direction. However, we notice that the above method has significant limitation that should be addressed. Before we introducing our method, we give some preliminary knowledge on RKHS embeddings.

3 Preliminary on Reproducing Kernel Hilbert Space Embeddings

Kernel methods [17] are a class of machine learning algorithms that map the data from the original space implicitly to a high dimensional or even infinite dimensional feature space \( \mathcal{H} \). One can get rid of computing the coordinates of the data in that space explicitly if the algorithm can reduce to inner products of feature vectors of all data points which can be easily calculated as the kernel function of any two data points. This is called the kernel trick [17]. The kernel function essentially act as a similarity function between a pair of data points and thus kernel methods are categorized as a typical method of instant-based learning.

\[
\phi(x) := k(\cdot, x),
\]

where \( \langle \phi(x), \phi(x') \rangle = \langle k(\cdot, x), k(\cdot, x') \rangle = k(x, x') \) and \( k(x, x') \) is a positive definite kernel. The kernel mean embedding [19] of a probability density \( p(x) \) is defined as:

\[
\mu_X = \int \phi(x)p(x)dx,
\]

One can simply interpret the kernel mean embedding as the vector of (higher order) moments. This interpretation is exactly true if one uses a polynomial kernel \( k(x, x') = (x^T x' + 1)^d \), where \( d > 0 \). It has been shown that if the kernel is characteristic [20], e.g. a Gaussian kernel, then the mapping a probability distribution to its kernel mean embedding is injective, i.e. we lose no information during the mapping. The conditional embeddings of the conditional distribution \( p(Y|X) \) is a sweep out a family of points in the RKHS [20], each one of which is essentially the kernel mean embedding of the conditional distribution \( p(Y|x) \) indexed by a fixed value of the conditioning variable \( x \). It is shown in [20] that under a mild assumption that \( E_{Y|X}[g(Y)] \in \mathcal{H}_X \), the conditional mean embedding can be obtained by Eq. [4]

\[
\mu_{Y|x} = C_{YX}C_{XX}^{-1}\phi(x),
\]

where \( C_{YX} := \int \phi(y) \otimes \phi(x)p(x, y)dx dy \) and \( C_{XX} := \int \phi(x) \otimes \phi(x)p(x)dx \). The empirical estimation of the kernel mean embedding and the conditional mean embedding given a set of observation \( X = [x_1, x_2, \cdots, x_n] \) and \( Y = [y_1, y_2, \cdots, y_n] \):

\[
\hat{\mu}_X = \frac{1}{n}\Phi 1, \quad \hat{\mu}_{Y|x} = \Psi (HK_x + \lambda n I)^{-1}k_x,
\]

where \( \Phi = [\phi(x_1), \phi(x_2), \cdots, \phi(x_n)] \), \( \Psi = [\phi(y_1), \phi(y_2), \cdots, \phi(y_n)] \), \( K_x \) is the kernel Gram matrix of \( x \), i.e. \( K_{ij} = k(x_i, x_j) \) and \( H = I - \frac{1}{n}11^T \) with \( I \) as an identity matrix and \( 1 \) is a vector of \( 1 \)s of appropriate dimension. The (conditional) kernel mean embeddings provide compact and nonparametric representation of the (conditional) distribution. Manipulations of the probability distribution such as complicated operations on probability distribution in Bayesian inference can easily reduces to matrix manipulation in the RKHS. For example, Maximum Mean Discrepancy (MMD) [5] was proposed for two sample test. A Kernel Bayes Rule (KBR) [4] was proposed to conduct Bayesian inference in the RKHS. Given that the RKHS embedding has solid theoretical support and is easy to use, it is adopted in our paper to measure the intrinsic invariance of the conditional distributions.

4 Causal Discovery by Intrinsic Invariance

Although positive results were reported on synthetic dataset and some real data in [11], there are some potential problems regarding the discriminative power of the RKHS norm-based method which is essentially calculating the variance of the conditional mean embedding norms. Back to the motivating example, we notice that in the anti-causal direction, the conditional density in red and the one in
We further assume \( \phi \) where \( \alpha \). According to Lemma 1, we see that even thought two probability densities are very different, e.g. for \( N \) Kernel Hilbert Space (RKHS) entailed by a posi-

tive definite kernel \( k \), we have \( \| \mu_p \|_{H_X} \) and \( \| \mu_q \|_{H_X} \) which correspond to the probability density function \( p(x) \) and \( q(x) = p(-x) \) are equal if a stationary (translation invariant) kernel is used.

**Lemma 1.** Denote the domain of \( x \) as \( X \), and if \( X \) is symmetric with respect to \( y \)-axis, given two probability densities \( p(x) \) and \( q(x) \) where \( q(x) = p(-x) \), we attain

\[
\| \mu_p \|_{H_X} = \| \mu_q \|_{H_X},
\]

where \( \mu \) is the kernel mean embedding with respect to a stationary kernel \( k(x, x') = k(x - x') \).

**Proof.** According to Bochner’s theorem [15], for a stationary kernel \( k(x - x') \), we have:

\[
\phi(x) = [\cos(\omega_1^T x), \sin(\omega_1^T x), \cdots, \cos(\omega_N^T x), \sin(\omega_N^T x)]^T,
\]

where \( N_X \) is the dimension of the feature space. We attain \( \mu_p = [\rho_1, \rho_2, \cdots, \rho_N, \xi_N] \), where \( \rho_i = \int \cos(\omega_i^T x)p(x)dx \) and \( \xi_i = \int \sin(\omega_i^T x)p(x)dx \) and thus \( \| \mu_p \|_{H} = \sum_{i=1}^{N} (\rho_i^2 + \xi_i^2) \).

Similarly, we have \( \mu_q = [\rho_1, -\rho_2, \cdots, -\rho_N, -\xi_N] \) and thus \( \| \mu_q \|_{H} = \sum_{i=1}^{N} (\rho_i^2 + \xi_i^2) \). Consequently, we show that \( \| \mu_p \|_{H_X} = \| \mu_q \|_{H_X} \).

According to Lemma 1, we see that even thought two probability densities are very different, e.g. for skewed distribution, \( p(x) \) and \( q(x) \) are different, but they share the same norm.

Similar conclusion can be drawn for more general cases and is justified in Lemma 2.

**Lemma 2.** Given an arbitrary probability density \( p(x) \in H_X \), where \( H_X \) is a Reproducing Kernel Hilbert Space (RKHS) entailed by a positive definite kernel \( k(x, x') \), then with high probability there exists at least one probability density \( q(x) \in H_X \) and \( q(x) \neq p(x) \) such that

\[
\| \mu_p \|_{H_X} = \| \mu_q \|_{H_X},
\]

where \( \mu_p \) and \( \mu_q \) are the kernel mean embeddings of \( p(x) \) and \( q(x) \) respectively.

**Proof.** Given a positive definite kernel \( k(x, x') \), according to Mercer’s Theorem [17], we have \( x \mapsto \phi(x) \), where \( \phi(x) = [\sqrt{\lambda_1} \phi_1(x), \sqrt{\lambda_2} \phi_2(x), \cdots, \sqrt{\lambda_N} \phi_N(x)] \), where \( \lambda_i > 0 \) and \( \int \phi_i(x) \phi_j(x)dx = \delta_{ij} \).

We further assume \( \phi_i(x) \) are integrable, \( \forall i \), then we write \( \theta_i = \int \sqrt{\lambda_i} \phi_i(x)dx \) and thus \( \theta = \int \phi(x)dx \). For an arbitrary probability density \( p(x) \in H_X \), we can represent it as:

\[
p(x) = 1/\alpha^T \theta \phi(x)^T \alpha,
\]

where \( \alpha \) is a vector of coefficient. By definition, the RKHS mean embedding of \( p(x) \) is obtained as

\[
\mu_p = \int \phi(x)p(x)dx = \frac{1}{\alpha^T \alpha} \Lambda \alpha,
\]

where \( \Lambda \) is a diagonal matrix with \( \Lambda_{ii} = \lambda_i \). The norm of \( \mu_p \) can be easily calculated as \( \| \mu_p \|_{H_X}^2 = 1/(\alpha^T \theta)^2 \alpha^T \Lambda^2 \alpha = 1/(\alpha^T \theta)^2 \sum_{i=1}^{N} \lambda_i^2 \alpha_i^2 \).
Now we construct another probability density \(q(x) = \frac{1}{\beta^T \theta} \phi(x)^T \beta\). Without loss of generality, we assume that \(\beta_i = \alpha_i, \forall i \geq 2\). Similarly, we have \(\|\mu_q\|^2_H = 1/(\beta^T \theta)^2 \beta^T \Lambda^2 \beta = 1/(\beta^T \theta)^2 \sum_{i=1}^{N_H} \lambda^2_i \beta^2_i\). In order to make \(\|\mu_p\|^2_H = \|\mu_q\|^2_H\), we construct \(\beta_1\) and \(\beta_2\) in the way that:

\[
\begin{align*}
\theta_1 \beta_1 + \theta_2 \beta_2 &= \theta_1 \alpha_1 + \theta_2 \alpha_2 = C_1 \\
\lambda^2_1 \beta^2_1 + \lambda^2_2 \beta^2_2 &= C^2_2,
\end{align*}
\]

where \(C_2 > 0\). We attain \((\lambda_1/C_2)^2 \beta^2_1 + (\lambda_2/C_2)^2 \beta^2_2 = 1\). Let \(\beta_1 = C_2/\lambda_1 \sin(\varphi)\) and \(\beta_2 = C_2/\lambda_2 \cos(\varphi)\), we obtain:

\[
\theta_1 C_2/\lambda_1 \sin(\varphi) + \theta_2 C_2/\lambda_2 \cos(\varphi) = C_1.
\]

In order to ensure a solution exists for Eq. 8, we need to prove that \(C_1/\sqrt{(\theta_1 C_2/\lambda_1)^2 + (\theta_2 C_2/\lambda_2)^2} \leq 1\). We show that

\[
\theta_1^2 C_2^2/\lambda_1^2 + \theta_2^2 C_2^2/\lambda_2^2 - C_1^2 = \theta_1^2/\lambda_1^2 (\lambda_2^2 \alpha_1^2 + \lambda_2^2 \alpha_2^2) + \theta_2^2/\lambda_2^2 (\lambda_1^2 \alpha_1^2 + \lambda_1^2 \alpha_2^2) - (\theta_1 \alpha_1 + \theta_2 \alpha_2)^2
\]

\[
= \lambda^2_1 \beta^2_1 \alpha_1^2 + \lambda^2_2 \beta^2_2 \alpha_2^2 - 2\theta_1 \theta_2 \alpha_1 \alpha_2 \geq 2|\theta_1 \theta_2 \alpha_1 \alpha_2| - 2\theta_1 \theta_2 \alpha_1 \alpha_2 \geq 0.
\]

Consequently, we prove that Eq. 8 holds and there exist two solutions, i.e. \(\varphi = \arcsin(C_1/\sqrt{(\theta_1 C_2/\lambda_1)^2 + (\theta_2 C_2/\lambda_2)^2} - \omega\) and \(\varphi = \pi - \arcsin(C_1/\sqrt{(\theta_1 C_2/\lambda_1)^2 + (\theta_2 C_2/\lambda_2)^2} - \omega), \) where \(\sin(\omega) = \theta_2/(\lambda_2 \sqrt{\beta_1^2/\lambda_1^2 + \beta_2^2/\lambda_2^2})\) and \(\cos(\omega) = \theta_1/(\lambda_1 \sqrt{\beta_1^2/\lambda_1^2 + \beta_2^2/\lambda_2^2}).\) Two solutions collapse to one if and only if \(C_1/\sqrt{(\theta_1 C_2/\lambda_1)^2 + (\theta_2 C_2/\lambda_2)^2} = 1\) which rarely happens as it requires mutual adjustment of the probability density function \(p(x)\) and the kernel function. □

The intuitive interpretation of the proof can also be elucidated in Fig. 2. The solution \((\beta_1, \beta_2)\) of the first equation \(\theta_1 \beta_1 + \theta_2 \beta_2 = C_1\) forms a line and the solution of the second equation \(\lambda^2_1 \beta^2_1 + \lambda^2_2 \beta^2_2 = C^2_2\) forms an ellipse and thus the solution of Eq. 7 is the intersection of the line and the ellipse. Note that the intersection should happen as \((\alpha_1, \alpha_2)\) is already a solution to Eq. 7. With high probability, there are two distinct intersection points as shown in Fig. 2 except for some rare cases that the points collapse to a single point when the line is the tangent line of the ellipse. This is rare because it requires mutual adjustment between \(\alpha, \theta, \beta, \lambda\) which in turn essentially requires the mutual adjustment between \(\phi(x)\) and \(p(x)\). According to Lemma 2, we see that the RKHS norm which is directly applied to the conditional distribution instantiated at different value is not discriminative enough. There are conditional distributions with significant distinction but they can have equal norms and thus it leads to some problems for the proposed KCDC algorithm in [11].

### 4.2 Causal Discovery via Kernel Intrinsic Invariance Measure

Realizing the limitation of the norm based approach, we propose our method which measures the norm of the difference of the kernel mean embeddings corresponding to conditional distributions instantiated at different values, instead of measuring the difference of their norms. However, a naive application of this idea might not work because even in the causal direction, conditional distributions instantiated at different cause values are not NOT identical. They could be different with each other in terms of the location and the scale, although we are more interested in higher order moments that are more relevant to the shape of the density function. For example, in a toy example proposed in [11], we have \(y = x^3 + x + \epsilon\), where \(\epsilon \sim \mathcal{N}(0, 1)\). Even in the causal direction, we get \(p(y|x) = \exp(-0.5(y - x^3 - x)^2)\). Conditional distributions instantiated at different \(x\) are all Gaussian distributions but they have different means and thus they are not identical, neither are their kernel mean embeddings and the corresponding norms (this can be easily verified if one use a polynomial kernel). However, the location and scale information of a distribution are not that interesting to us when it comes to causal discovery as we are more keen on the higher order statistics that reflect shape information.

This observation motivates our method to capture more intrinsic information of the probability density function. Mathematically, we define the following score that captures the “intrinsic” variation of the conditional distribution instantiated at different values of the conditioning variable \(x\) or \(y\) for two
hypothetic directions. Without loss of generality, we show definition of the score in the direction of \( x \rightarrow y \):

\[
S_{x \rightarrow y} := \min_W \frac{1}{n} \sum_{i=1}^{n} \| W^T \mu_{y|x} - \frac{1}{n} \sum_{j=1}^{n} W^T \mu_{y|x} \|_{H_y}^2.
\]

(9)

The interpretation of Eq. (9) is that we calculate the norm of the difference of conditional distributions at different values. The score is zero if and only if all conditional distributions are the same according to the injectiveness of the kernel mean embedding. The matrix \( W \) is introduced to find the subspace which removes the effect of some trivial deviance like location and scale. Empirically, we can calculate the kernel embedding of the conditional distribution instantiated at different \( \tilde{x} \) as

\[
\mu_{y|x} = \Psi \begin{bmatrix} K_x + \lambda I \end{bmatrix}^{-1} k_{x_i},
\]

where \( \Psi = [\psi(y_1), \cdots, \psi(y_n)] \), \( K_x \) is the kernel gram matrix of \( x \) and \( k_{x_i} = k(x, x_i) \). Note that the solution of the above optimization problem lies in the span of \( \Psi \), and thus we can represent \( W \) by a linear combination of \( \psi(y_i) \), i.e. \( W = \Psi \tilde{W} \), where \( \tilde{W} \in \mathbb{R}^{n \times p} \) is the coefficient matrix. Consequently, we attain

\[
S_{x \rightarrow y} := \min_W \frac{1}{n} \sum_{i=1}^{n} \| W^T \tilde{K}_y (K_x + \lambda I)^{-1} k_{x_i} - \frac{1}{n} \sum_{j=1}^{n} \tilde{W}^T \tilde{K}_y (K_x + \lambda I)^{-1} k_{x_j} \|_{H_y}^2
\]

\[
= \min_W \frac{1}{n} \text{Tr} \left( W^T \tilde{K}_y (K_x + \lambda I)^{-1} K_x H \tilde{K}_y (K_x + \lambda I)^{-1} \tilde{K}_y \tilde{W} \right)
\]

(10)

To avoid trivial solution \( \tilde{W} = 0 \), we pose the constraint that \( \tilde{W}^T \tilde{W} = I \). Consequently, we have

\[
S_{x \rightarrow y} = \min_W \frac{1}{n} \text{Tr} \left( W^T \tilde{K}_y (K_x + \lambda I)^{-1} K_x H \tilde{K}_y (K_x + \lambda I)^{-1} \tilde{K}_y \tilde{W} \right) \quad \text{s.t.} \tilde{W}^T \tilde{W} = I.
\]

(11)

The intuitive interpretation of the proposed method is that we use the projection matrix \( W \) to find intrinsic deviance of the conditional mean embedding. The intrinsic deviance captures higher order statistics which is more relevant to the shape of the probability distribution function while discards the some less important information, e.g. the location and scale of a distribution. As an illustrative example, suppose we use a polynomial kernel \( k(x, x') = (x x' + 1)^d \), it can be easily shown that we essentially map \( x \) to a space with polynomials of the feature, i.e. \( x \mapsto [1, x, x^2, \cdots, x^d] \). Therefore, the kernel mean embedding of a distribution is in fact a vector of moments up to degree \( d \), i.e.

\[
\mu_x := \int \phi(x) p(x) dx = [1, m_1, m_2, \cdots, m_d]^T,
\]

where \( m_i \) denotes the \( i \)-th order moment. If the conditional distributions only differ from each other with mean and standard deviation (the first and second order moments), then the projection matrix \( W \) is expected to find the subspace that contains only higher order moments.

However, how to decide the rank of \( W \) is an open question. In this paper, we propose a simple but effective algorithm to choose the right rank, see Alg. 1. The basic idea is to project of the subspace which removes the effect of some trivial deviance like location and scale. Empirically, we use the projection matrix \( W \) to find the eigenvectors which preserve at least 90% of the energy of the whole spectrum.

### 4.3 Robust Kernel Intrinsic Invariance Measure by Importance Reweighting

Real world data is usually contaminated with noise and outliers. The estimation of the kernel mean embedding might be significantly biased due to the outliers in data. Furthermore, when estimating the conditional mean embedding, we want to eliminate any effect of the marginal distribution of the hypothetic cause due to finite sample size. We adopt an importance reweighting scheme as follows:

\[
C_{Y|X}^{ref} := \int \phi(y) \otimes \phi(x) p(y|x) u(x) dx, \quad C_{X|X}^{ref} := \int \phi(x) \otimes \phi(x) u(x) dx,
\]

(12)

and

\[
\mu_{Y|X}^{ref} = C_{Y|X}^{ref} \left( C_{X|X}^{ref} \right)^{-1} \phi(x), \quad \text{where} \quad u(x) \text{ is a reference distribution.}
\]

The empirical estimation is then obtained by

\[
\hat{\mu}_{Y|X}^{ref} = \Psi H R^{1/2} (H R^{1/2} K_x R^{1/2} H + \lambda n I)^{-1} R^{1/2} H k_{X},
\]

(13)
Algorithm 1 Framework of the Kernel Intrinsic Invariance Measure (KIIM)

Require:
The set of samples \( X = [x_1, x_2, \cdots, x_n], Y = [y_1, y_2, \cdots, y_n] \);
The regularization hyperparameter \( \lambda \);

Ensure:
The inferred causal direction, \( S_x \rightarrow y \), and \( S_y \rightarrow x \).

1: Compute Kernel Gram matrix \( K_y, K_x \) and the centering matrix \( H \);
2: Compute the Kernel Intrinsic Invariance Matrix \( M^{(1)} = K_y(K_x + \lambda I)^{-1} K_x H K_x (K_x + \lambda I)^{-1} K_y \) for the hypothetic direction \( x \rightarrow y \);
3: Conduct eigen-decomposition of \( M^{(1)} = U^{(1)} \Pi^{(1)} (U^{(1)})^T \) and calculate \( S_x \rightarrow y = \sum_{i=k_1}^n \pi_i^{(1)} \),
   where \( \pi_i \), \( \forall i \geq k_1 \) are the bottom-\( k \) smallest eigenvalues such that \( \sum_{i=k_1}^n \pi_i^{(1)} / \sum_{i=1}^n \pi_i^{(1)} \geq 0.9 \) and \( k_1 \) is the largest number when the inequality holds.
4: Repeat Step 3 for the other direction.
5: The causal direction is inferred as \( x \rightarrow y \) if \( S_x \rightarrow y < S_y \rightarrow x \) or \( y \rightarrow x \) if \( S_x \rightarrow y > S_y \rightarrow x \). No conclusion will be made if \( S_x \rightarrow y = S_y \rightarrow x \).
6: return The causal direction, \( S_x \rightarrow y, S_y \rightarrow x \);

where \( R \) is a diagonal reweighting matrix with \( |R|_{ij} = u(x_i) / p(x_i) \). The main body of the algorithm does not change except for the calculation of \( M^{(1)} \) and \( M^{(2)} \) in Alg. 1. We name this variant of our algorithm Rw-KIIM meaning Reweighted Kernel Intrinsic Invariance Measure.

5 Experiment

In this section, we conduct experiments using both synthetic data and a real world dataset called TCEP (Tubigen Causal Effect Pairs). We compare our methods with some state-of-the-art methods including KCDC\(^1\), IGCI\(^2\), ANM\(^3\) and LiNGAM\(^4\). For IGCI, we use the entropy-based methods with two different reference distributions (Gaussian and Uniform distribution). We use \( \lambda = 3 \) for the regularization hyperparameter when calculating the conditional mean embedding in Eq.5.

In the following experiment, we use the composite kernel for KIIM which is the multiplication of the RBF kernel \( k(x, x') = \exp(-||x-x'||^2 / \sigma^2) \) with median heuristic for kernel width and a log kernel \( k(x, x') = -\log(||x-x'||^2 + 1) \) and a rational quadratic kernel \( k(x, x') = 1 - \frac{||x-x'||^2}{||x-x'||^2 + 1} \).

For KCDC, we use log kernel for the input and rational quadratic kernel for the output as in \( 11 \).

Table 1: Performance of synthetic dataset

|            | ANM-1     | KCDC      | KIIM      | Rw-KIIM    | IGCI(Gaussian) | IGCI(Uniform)     | ANM     |
|------------|-----------|-----------|-----------|------------|----------------|--------------------|---------|
| Gaussian   | 94.0% ± 2.6% | 100.0% ± 0.0% | 100.0% ± 0.0% | 98.1% ± 4.1% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% |
| Uniform    | 61.1% ± 3.9% | 100.0% ± 0.0% | 100.0% ± 0.0% | 99.4% ± 0.97% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% |
|            | 59.8% ± 4.3% | 89.8% ± 2.4% | 98.2% ± 2.1% | 44.8% ± 4.0% | 93.6% ± 4.1% | 70.1% ± 4.8% | 67.6% ± 3.2% |

|            | ANM-2     | KCDC      | KIIM      | Rw-KIIM    | IGCI(Gaussian) | IGCI(Uniform)     | ANM     |
|------------|-----------|-----------|-----------|------------|----------------|--------------------|---------|
| Gaussian   | 23.3% ± 2.1% | 100.0% ± 0.0% | 100.0% ± 0.0% | 99.2% ± 2.1% | 100.0% ± 0.0% | 100.0% ± 0.0% | 99.9% ± 3.3% |
| Uniform    | 24.6% ± 4.1% | 100.0% ± 0.0% | 100.0% ± 0.0% | 99.9% ± 3.3% | 100.0% ± 0.0% | 100.0% ± 0.0% | 0.2% ± 0.0% |
|            | 56.8% ± 4.5% | 57.0% ± 4.2% | 49.4% ± 5.6% | 49.4% ± 6.1% | 49.4% ± 6.1% | 67.6% ± 3.2% | 67.6% ± 3.2% |

|            | MNM-1     | KCDC      | KIIM      | Rw-KIIM    | IGCI(Gaussian) | IGCI(Uniform)     | ANM     |
|------------|-----------|-----------|-----------|------------|----------------|--------------------|---------|
| Gaussian   | 99.2% ± 8.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 1.0% ± 1.3% |
| Uniform    | 97.9% ± 1.1% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 28.4% ± 5.7% | 28.4% ± 5.7% |
|            | Complex   | KCDC      | KIIM      | Rw-KIIM    | IGCI(Gaussian) | IGCI(Uniform)     | ANM     |
| Gaussian   | 27.6% ± 5.8% | 99.8% ± 0.4% | 99.8% ± 0.4% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 6.9% ± 1.9% |
| Uniform    | 5.1% ± 1.4% | 94.1% ± 2.0% | 91.7% ± 2.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 100.0% ± 0.0% | 15.1% ± 3.8% |

5.1 Synthetic Data

In this section, we evaluate pairwise causal discover algorithms on data generated by 5 different data generation mechanisms following \( 11 \). Details of these mechanisms are given as follows. ANM-1: \( y = x^3 + x + \epsilon \); ANM-2: \( y = x + \epsilon \); MNM-1: \( y = (x^3 + x) \exp(\epsilon) \); MNM-2: \( 1 \) Although positive results were reported in \( 11 \), unfortunately we are not able to reproduce the results reported in the paper.

\(^2\) https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox
\[
y = (\sin(10x) + \exp(3x)) \exp(\epsilon) \quad \text{and} \quad \text{CNM: } y = (\log(x + 10) + x^2)\epsilon
\]
and the noise distribution is specified in Tab[1]. We generate 100 samples from each data generation mechanism for different algorithm to infer the causal direction. Experiments are conducted for 100 independent trials and results of different algorithms are reported in Tab[1]. We observe that IGCI is quite robust and ANM performs well when the data generation mechanism is indeed additive noise model. Unfortunately, we are not able to reproduce positive results reported in [11]. In this paper, we are using a direct version of KCDC without majority vote and the confident measure [11] because these extra processes are not used in our algorithm IKKM. Even without this extra process, IKKM and Rw-KIIM perform quite well except for the linear ANM with uniform noise.

In order to justify the necessity of using a projection matrix \( W \) to a lower dimensional space, we compare the performance of IKKM with different ranks of \( W \). This results in using the algorithm exploiting eigenvalues ranging from the whole spectrum to only the smallest one. Two mechanisms are used in this experiments as shown in Fig.[5a]. Interestingly, we find that the algorithm using the whole spectrum does not perform the best but the one discarding the top-1 eigenvalue performs consistently the best. This result justifies our motivation and argument: we need intrinsic deviance/invariance measurement that captures only higher order statistics of the shape of the conditional distribution. Trivial difference arising from the location and the scale might not be beneficial or even harmful for causal discovery.

### 5.2 Tuebingen Cause-Effect Pairs (TCEP)

In this section, we verify the performance of our algorithm on real world data. We use the open benchmark called Tuebingen Cause-Effect Pair[3], which has been widely used to evaluate causal discovery algorithms. The whole dataset contains 108 cause-effect pairs taken from 37 different data sets from various domains [12] with known ground truth. We do not use some pairs as both \( x \) and \( y \) are high-dimensional variables in pair 52, 53, 54, 55, 71, 105 and there are missing values in pairs 81, 82 and 83. The ground truth direction in pair 86 is not mentioned in the data description document and thus is not used in our experiments. Fig[3b] shows our proposed algorithm outperform the state-of-the-art methods significantly.

### 6 Conclusion

In this paper, we focus on causal discovery for cause-effect pairs along the direction of Independent Mechanism (IM) principle. We prove that the existing norm-based state-of-the-art which only compare the norms of conditional mean embedding might lose discriminative power. To solve this problem, we propose a Kernel Intrinsic Invariance Measure (KIIM) to capture the intrinsic invariance of the conditional distribution, i.e. the higher order statistics corresponding to the shapes of the density functions. Experiments with synthetic data and real data justify the effectiveness of our proposed algorithm and supports our argument that we indeed need to look for higher order statistics/intrinsic invariance for causal discovery.

[3]https://webdav.tuebingen.mpg.de/cause-effect/
References

[1] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. arXiv preprint arXiv:1901.10912, 2019.

[2] K. Budhathoki and J. Vreeken. Mdl for causal inference on discrete data. In 2017 IEEE International Conference on Data Mining (ICDM), pages 751–756. IEEE, 2017.

[3] Z. Chen, K. Zhang, L. Chan, and B. Schölkopf. Causal discovery via reproducing kernel hilbert space embeddings. Neural computation, 26(7):1484–1517, 2014.

[4] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes’ rule: Bayesian inference with positive definite kernels. The Journal of Machine Learning Research, 14(1):3753–3783, 2013.

[5] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In Advances in neural information processing systems, pages 585–592, 2008.

[6] C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen. Causal structure learning. Annual Review of Statistics and Its Application, 5:371–391, 2018.

[7] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In Advances in neural information processing systems, pages 689–696, 2009.

[8] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. Artificial Intelligence, 182:1–31, 2012.

[9] D. Janzing and B. Scholkopf. Causal inference using the algorithmic markov condition. IEEE Transactions on Information Theory, 56(10):5168–5194, 2010.

[10] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou. Discovering causal signals in images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6979–6987, 2017.

[11] J. Mitrovic, D. Sejdinovic, and Y. W. Teh. Causal inference via kernel deviance measures. In Advances in Neural Information Processing Systems, pages 6986–6994, 2018.

[12] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17(1):1103–1204, 2016.

[13] J. Pearl. Causality: models, reasoning, and inference. Econometric Theory, 19(675-685):46, 2003.

[14] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5):947–1012, 2016.

[15] W. Rudin. Fourier analysis on groups, volume 121967. Wiley Online Library, 1962.

[16] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. Semi-supervised learning in causal and anticausal settings. In Empirical Inference, pages 129–141. Springer, 2013.

[17] B. Scholkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

[18] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7(Oct):2003–2030, 2006.

[19] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In International Conference on Algorithmic Learning Theory, pages 13–31. Springer, 2007.

[20] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. IEEE Signal Processing Magazine, 30(4):98–111, 2013.

[21] P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. In Applied informatics, volume 3, page 3. SpringerOpen, 2016.

[22] X. Sun, D. Janzing, and B. Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. Neurocomputing, 71(7-9):1248–1256, 2008.