PHSIC against Random Consistency and Its Application in Causal Inference

Jue Li , **Yuhua Qian*** , **Jieting Wang** and **Saixiong Liu**
Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China
lijue688@163.com, jinchengqyh@126.com, jietingwang@email.sxu.edu.cn, liu_saixiong@126.com

Abstract

The Hilbert-Schmidt Independence Criterion (HSIC) based on kernel functions is capable of detecting nonlinear dependencies between variables, making it a common method for association relationship mining. However, in situations with small samples, high dimensions, or noisy data, it may generate spurious associations, causing two unrelated variables to have certain scores. To address this issue, we propose a novel criterion, named as Pure Hilbert-Schmidt Independence Criterion (PHSIC). PHSIC is achieved by subtracting the mean HSIC obtained under random conditions from the original HSIC value. We demonstrate three significant advantages of PHSIC through theoretical and simulation experiments: (1) PHSIC has a baseline of zero, enhancing the interpretability of HSIC. (2) Compared to HSIC, PHSIC exhibits lower bias. (3) PHSIC enables a fairer comparison across different samples and dimensions. To validate the effectiveness of PHSIC, we apply it to multiple causal inference tasks to measure the independence between cause and residual. Experimental results demonstrate that the causal model based on PHSIC performs well compared to other methods in scenarios involving small sample sizes and noisy data, both in real and simulated datasets.

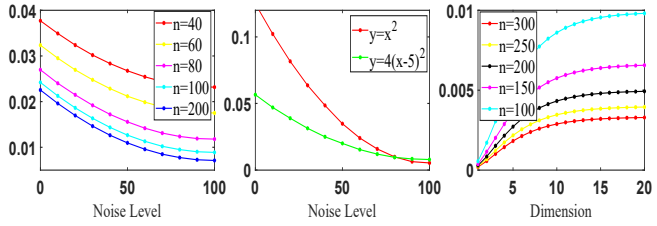
1 Introduction

Association analysis is the basis of data mining and machine learning, which has been applied in many fields. Examples include feature selection [Li *et al.*, 2021; Song *et al.*, 2007b], gene analysis [Yamanishi *et al.*, 2004], causal inference [Hoyer *et al.*, 2008], etc. There are various forms of correlation measurement, which can be roughly classified into two categories: the construction method based on the reduction error ratio [Puth *et al.*, 2015; Hotelling, 1992] and the construction method based on the independence test [Jaworski *et al.*, 2010; Schweizer and Wolff, 1981]. The typical representative of the first construction method is the person correlation coefficient, but it can only recognize simple linear relationships. The typical representatives of the second construction method are mutual information and HSIC. Although both mutual information and HSIC can identify different forms of dependency relationships, the empirical estimation of mutual information is expensive, while the empirical estimate of HSIC is much simpler-just the trace of the product of gram matrices.

The good nature of HSIC drives the development of many fields. For example, in terms of neural network optimization, since the maximum information of continuous variables is difficult to calculate, Ma et al. [Ma *et al.*, 2020] used the Hilbert-Schmidt information bottleneck (HSIC Information Bottleneck) as the regular term in the neural network optimization objective to improve the adversarial robustness of the model. In clustering problems, HSIC is introduced many times to measure the independence of cluster centers [Niu *et al.*, 2010; Song *et al.*, 2007a]. In the field of feature selection [Liaghat and Mansoori, 2019; Song *et al.*, 2007b], HSIC is used to calculate the degree of association between features and labels. In the field of causal inference [Hoyer *et al.*, 2008], HSIC is used to measure the independence between independent variables and residuals in additive noise models, so as to judge the causal direction. These are attributed to its three good characteristics: (1) it has smaller bias than other indicators; (2) it can capture many complex nonlinear dependencies without explicitly considering random variables ; (3) its empirical estimate is the trace of the product of gram matrices, which makes the objective function easy to solve.

With sufficient sample size, HSIC undoubtedly shows good performance. However, in some application scenarios, data collection is difficult due to some complex factors, so the collected samples are relatively sparse. At the same time, this random error will be amplified in high-dimensional and noisy data. HSIC generates random errors in limited samples, which causes a series of problems. (1) Firstly, as shown in Figure 1(a), in an independent situation, there is no 0 baseline, which lacks interpretability. And we can see that the smaller the sample, the higher the score of HSIC, and there is sample bias. (2) Secondly, as shown in Figure 1(b), due to the bias of limited samples, random consistency can also occur when comparing different correlations, which can seriously affect the performance of the application task. (3) Finally, as shown in Figure 1(c), in independent situations, as the dimen-

*The corresponding author



(a) (b) (c)

Figure 1: Disadvantages of HSIC. (a) The variation of HSIC values for variables x and y constructed by the y=x function under different independent uniform noise and sample sizes. (b) The values of HSIC for x and y under different proportions of independent uniform noise using variables constructed by $y = x^2$ and $y = 4(x-5)^2$. (c) The values of HSIC in different dimensions when variables x and y are independent.

sion gets higher, the value of HSIC gets larger, which leads to the dimension deviation problem.

Not only does HSIC exhibit random consistency, but random consistency is a common problem in statistics and machine learning. In classification tasks, due to the limited number of samples and the presence of noise, the results generate random consistency, which makes machine learning algorithms lack generalization ability. In order to alleviate the random consistency in classification tasks, Wang et al. [Wang *et al.*, 2022] proposed a pure consistency index, which verified the low bias, learning substitutability, and high recognizability of pure accuracy frameworks compared to accuracy. And Romano et al. [Romano *et al.*, 2016] improved the classification performance of decision trees by eliminating random factors in the Gini index. In clustering tasks, the results may have biases due to factors such as the number of clusters and sample size. Vinh et al.[Vinh *et al.*, 2009] emphasized the importance of eliminating mutual information random consistency in clustering communities, especially when the data size is relatively small relative to the number of clusters.

Based on the above observations, we propose a pure HSIC framework that can alleviate the drawbacks of HSIC and ensure that the original HSIC properties remain unchanged. PHSIC is obtained by separating the mean of random parts from the original HSIC values. To emphasize the advantages of PHSIC, we have demonstrated theoretically and experimentally that PHSIC has smaller deviations compared to HSIC, especially in independent cases where PHSIC has an error order of m^{-2} and HSIC has an error order of m^{-1}, where m represents the number of samples. At the same time, it has also been proven that the random part is the reason for the deviation of HSIC in dimensions, samples, and noise.

Association is an important issue in causal discovery. Current causal inference methods can generally be summarized as 'association + hypothesis'. For example, in constraint-based causal inference methods [Spirtes *et al.*, 2000; Ogarrio *et al.*, 2016], these methods mine the causal relationship between variables through the conditional independence test under the Markov faithful assumption. In function-based causal inference methods [Shimizu *et al.*, 2006; Zhang and Hyvarinen, 2012], under the assumption of independent causal mechanism, the causal direction is further inferred by measuring the correlation between residuals and causes through correlation indicators. It can be seen that a robust correlation indicator is the premise to promote the performance of causal inference methods. Currently, identifying causal directions is difficult in small sample and noisy scenarios. Therefore, we applied PHSIC to the classical causal models to further improve the performance of causal inference model. The main code and supplementary material have been made available online[1]. The contributions of this paper are as follows:

- We analyze the random consistency phenomenon of HSIC from both theoretical and experimental perspectives.
- We propose a PHSIC framework to eliminate random consistency, proving theoretically and experimentally that PHSIC has smaller deviations than HSIC, and discussing the properties of PHSIC.
- We apply PHSIC to multiple classic causal inference tasks and experimentally verify the robustness of PHSIC based causal models to small sample and noisy data.

2 Hilbert-Schmidt Independence Criterion

In this section, we briefly introduce HSIC and demonstrate the bias of $HSIC_b(Z)$.

Suppose X and Y are two random variables. Their values belong to χ and γ respectively. $\mathcal{F}$ presents reproducing Hilbert space for each $x \in X$. We define that $\phi_X(x) \in \mathcal{F}$ as an element in a reproducing Hilbert space corresponding to each point $x \in \chi$. Assuming the existence of a continuously bounded positive definite kernel $k_X : \chi^2 \to R$, where

$$k_X(x, x') =< \phi_X(x), \phi_X(x') > . \tag{1}$$

For example, the Gaussian kernel $k(x_1, x_2) = exp(-||x_1 - x_2||^2/\sigma^2)$ is the primary choice in many literature. It is a continuously bounded kernel function, and we mainly default it to the selected kernel function in this article. Similarly, for a random variable Y, let $\mathcal{G}$ be a reproducing kernel Hilbert space with kernel $k_Y : \gamma^2 \to R$ and feature mapping $\phi_Y : \gamma \to \mathcal{G}$.

In order to observe the correlation between X and Y, Gretton et al. [Gretton *et al.*, 2005] defined cross covariance operator,

$$C_{XY} = \mathbb{E}_{XY}[(\phi_X - \mu_X) \otimes (\phi_Y - \mu_Y)], \tag{2}$$

C_{XY} means a linear operation from $\mathcal{G} \to \mathcal{F}$. And $\otimes$ represents the tensor product, $\mu_X = \mathbb{E}_X(\phi_X)$, $\mu_Y = \mathbb{E}_Y(\phi_Y)$ represents the average value of the corresponding elements of random variables X and Y in a reproducing Hilbert space, respectively. This equation cannot quantify correlation, so the square of Hilbert normal form is used instead of the cross covariance operator and it can be expressed in terms of kernels:

$$\begin{aligned} &HSIC(\mathcal{F}, \mathcal{G}, P_{X,Y}) = ||C_{XY}||^2_{HS}, \\ &= \mathbb{E}_{xx'yy'}[k(x, x')l(y, y')] + \mathbb{E}_{xx'}[k(x, x')]\mathbb{E}_{yy'}[l(y, y')] \\ &- 2\mathbb{E}_{xy}[\mathbb{E}_{x'}[k(x, x')]\mathbb{E}_{y'}[l(y, y')]], \end{aligned} \tag{3}$$

[1]https://github.com/lijue688/main.git

which is a measure of the statistical dependence of X and Y. When random variables X and Y are independent then $HSIC(\mathcal{F}, \mathcal{G}, P_{X,Y}) = 0$. If X and Y are correlated, then $HSIC(\mathcal{F}, \mathcal{G}, P_{X,Y}) > 0$. And as the values of X and Y become more correlated, the value of HSIC becomes larger.

Gretton et al. [Gretton *et al.*, 2007] define the following empirical HSIC estimator for an i.i.d. sample $Z = \{(x^i, y^i)\}_{i=1,\dots,m}$:

$$\begin{aligned} HSIC_b(Z) &= \frac{1}{m^2}\sum_{i,j}^{m} k_{ij}l_{ij} + \frac{1}{m^4}\sum_{i,j,q,r}^{m} k_{ij}l_{qr} \\ &- 2\frac{1}{m^3}\sum_{i,j,q}^{m} k_{ij}l_{iq} = \frac{1}{m^2}trace(KHLH), \end{aligned} \tag{4}$$

where m is the number of data points, K is the $m \times m$ kernel matrix for X and L is that for Y:

$$K_{i,j} = k_X(x^i, x^j), L_{i,j} = l_Y(y^i, y^j), \tag{5}$$

and H is the $m \times m$ matrix defined by:

$$H := \mathrm{I} - \frac{1}{m} I \cdot I^T, \tag{6}$$

where I is an identity matrix of $m \times m$ and I is a vector of $m \times 1$.

Gretton et al. show that for $m \to \infty$, $HSIC_b(Z) \to 0$ if and only if X is independent of Y [Gretton *et al.*, 2007]. Next, we explored the inherent bias of $HSIC_b(Z)$.

Theorem 1 (Bias of $HSIC_b(Z)$). *Let $\mathbb{E}(Z)$ denote the expectation over m independent copies(x_i, y_i) drawn from P_{xy}, then we have*

$$HSIC(P_{xy}, \mathcal{F}, \mathcal{G}) = \mathbb{E}_Z[HSIC_b(Z, \mathcal{F}, \mathcal{G})] - O(m^{-1}). \tag{7}$$

We provide the proof of this theorem in the appendix. It can be observed that although the empirical estimates of HSIC proposed by Gretton et al. [Gretton *et al.*, 2005; Gretton *et al.*, 2007] in 2005 and 2007 have different coefficients, one is $1/(m-1)^2$, and the other is $1/(m)^2$, their deviation orders are both $O(m^{-1})$. From its deviation order, it can be seen that when the sample size is relatively small, it will have a certain impact on the performance of HSIC.

3 Pure HSIC to Alleviate Random Consistency

In order to alleviate the random factors generated by $HSIC_b$, we first propose a framework for eliminating random consistency, and then explore the properties of PHSIC.

Definition 1 (Eliminating Random Consistency Framework). Due to the degree of consistency between some random variables being caused by their random distribution, we introduce a framework $PCM(Z_1, Z_2)$ for eliminating random consistency:

$$PCM(Z_1, Z_2) = CM(Z_1, Z_2) - RCM(Z_1, Z_2), \tag{8}$$

where Z_1, Z_2 are random variables. $CM(Z_1, Z_2)$ represents the degree of consistency between Z_1 and Z_2, which usually represents the degree of correlation or similarity between two variables. And $RCM(Z_1, Z_2)$ denotes the degree of random consistency between Z_1 and Z_2, which causes by factors such as limited number of samples, high dimensionas, and uniform noise. Assuming a set Z_1' is a uniform random variable with the same value and distribution as Z_1, the expected values of random consistency in random variables Z_1' and Z_2 can be used to calculate $RCM(Z_1, Z_2)$. Therefore, PCM will characterize pure consistency by the difference between the degree of consistency and the degree of random consistency.

Definition 2 (Empirical PHSIC). With the framework of equation (8), we define PHSIC:

$$\begin{aligned} &PHSIC(Z, \mathcal{F}, \mathcal{G}) \\ &= HSIC_b(Z, \mathcal{F}, \mathcal{G}) - \mathbb{E}(HSIC_0(Z, \mathcal{F}, \mathcal{G})), \end{aligned} \tag{9}$$

where $\mathbb{E}(HSIC_0(Z, \mathcal{F}, \mathcal{G}))$ represents its expected value under the null distribution. The analysis formula is shown in equation (10), which can obtain from [Gretton *et al.*, 2007].

$$\begin{aligned} \mathbb{E}(HSIC_0(Z, \mathcal{F}, \mathcal{G})) &= \frac{1}{m}(1 + ||u_x||^2||u_y||^2 \\ &- ||u_x||^2 - ||u_y||^2). \end{aligned} \tag{10}$$

Theorem 2 (Bias of PHSIC(Z)). *The bias of empirical PHSIC(Z) is as follows:*

$$\begin{aligned} &\mathbb{E}_Z[PHSIC(Z, \mathcal{F}, \mathcal{G})] - HSIC(P_{xy}, \mathcal{F}, \mathcal{G}) = \\ &\frac{1}{m}(-3\mathbb{E}_{xx'yy'}kl + 10\mathbb{E}_{xx'yy"}kl - 7\mathbb{E}_{xx'}k\mathbb{E}_{yy'}l) + O(m^{-2}). \end{aligned} \tag{11}$$

Proof. Gretton et al. proposes that the unbiased estimator of HSIC can be replaced by three u-statistics [Gretton *et al.*, 2007]:

$$\begin{aligned} HSIC(Z) &= \frac{1}{(m)_2}\sum_{(i,j)\in i_2^m} K_{ij}L_{ij} \\ &+ \frac{1}{(m)_4}\sum_{(i,j,q,r)\in i_4^m} K_{ij}L_{qr} - 2\frac{1}{(m)_3}\sum_{(i,j,q)\in i_3^m} K_{ij}L_{iq}, \end{aligned} \tag{12}$$

set i_n^m indicate that there are no duplicate combinations, the empirical averages can be used instead of expected values:

$$\mathbb{E}_Z[HSIC(Z, \mathcal{F}, \mathcal{G})] = HSIC(P_{xy}, \mathcal{F}, \mathcal{G}), \tag{13}$$

Same idea as Gretton et al. [Gretton *et al.*, 2007], we can obtain deviation between biased estimation of $HSIC_b(Z)$ and unbiased estimation of $HSIC(Z)$ through the three difference terms, as shown in equation (14), (15), and (16):

$$\begin{aligned} &\frac{1}{m^2}\sum_{i,j} k_{ij}l_{ij} - \frac{1}{(m)_2}\sum_{(i,j)\in i_2^m} k_{ij}l_{ij} \\ &= \frac{1}{m^2}\sum_{i} k_{ii}l_{ii} - \frac{1}{m(m)_2}\sum_{i} k_{ij}l_{ij}, \end{aligned} \tag{14}$$

$$\frac{1}{m^3}\sum_{i,j,q} k_{ij}l_{iq} - \frac{1}{(m)_3}\sum_{(i,j,q)\in i_3^m} k_{ij}l_{iq}$$
$$= \frac{1}{m^3}\sum_{(i,j)\in i_2^m}(k_{ii}l_{ij} + k_{ij}l_{ii} + k_{ij}l_{ij}) \quad (15)$$
$$- \frac{3}{m(m)_3}\sum_{(i,j,q)\in i_3^m} k_{ij}l_{iq} + O(m^{-2}),$$

$$\frac{1}{m^4}\sum_{i,j,q,r} k_{ij}l_{qr} - \frac{1}{(m)_4}\sum_{(i,j,q,r)\in i_4^m} k_{ij}l_{qr}$$
$$= \frac{1}{m^4}\sum_{(i,j)\in i_3^m}(k_{ii}l_{jq} + 4k_{ij}l_{iq} + k_{ij}l_{qq}) \quad (16)$$
$$- \frac{6}{m(m)_4}\sum_{(i,j,q,r)\in i_4^m} k_{ij}l_{qr} + O(m^{-2}).$$

Combining these three items, then seeking expectations on both sides, the deviation of $HSIC_b(Z)$ is:

$$bias(HSIC_b) = \mathbb{E}(HSIC_b - HSIC) = \frac{1}{m}(\mathbb{E}_{xy}kl$$
$$- 2\mathbb{E}_{xyy'}kl - 2\mathbb{E}_{xx'y}kl + \mathbb{E}_{xy'y"}kl + \mathbb{E}_{xx'y"}kl$$
$$- 3\mathbb{E}_{xx'yy'}kl + 10\mathbb{E}_{xx'yy"}kl - 6\mathbb{E}_{xx'}k\mathbb{E}_{yy'}l) + O(m^{-2}). \quad (17)$$

In the independent case,$-3\mathbb{E}_{xx'yy'}kl + 10\mathbb{E}_{xx'yy"}kl - 6\mathbb{E}_{xx'}kE_{yy'}l$ can be merged into one item. Ignoring the bias of $O(m^{-2})$, [Gretton *et al.*, 2007] derived the mean of HSIC under the null hypothesis, as shown in equation (18):

$$\mathbb{E}(HSIC_b - HSIC) = \frac{1}{m}(\mathbb{E}_{xy}kl + ||u_x||^2||u_y||^2$$
$$- \mathbb{E}_y l||u_x||^2 - \mathbb{E}_x k||u_y||^2) \quad (18)$$
$$= \frac{1}{m}TrC_{xx}TrC_{yy} = \mathbb{E}(HSIC_0).$$

Therefore, based on equations(13), (17) and (18), we can conclude:

$$\mathbb{E}_Z[HSIC_b(Z)] - \mathbb{E}(HSIC_0(Z)) - HSIC(P_{xy}, \mathcal{F}, \mathcal{G})$$
$$= \mathbb{E}_Z[PHSIC(Z)] - HSIC(P_{xy}, \mathcal{F}, \mathcal{G}) = \frac{1}{m}($$
$$- 3\mathbb{E}_{xx'yy'}kl + 10\mathbb{E}_{xx'yy"}kl - 7\mathbb{E}_{xx'}k\mathbb{E}_{yy'}l) + O(m^{-2}). \quad (19)$$

□

Corollary 1. *The deviation of $PHSIC(Z, \mathcal{F}, \mathcal{G})$ in independent cases is $O(m^{-2})$.*

Proof. Theorem 2 has already mentioned that in the independent case, $-3\mathbb{E}_{xx'yy'}kl + 10\mathbb{E}_{xx'yy"}kl - 6\mathbb{E}_{xx'}kE_{yy'}l$ can be merged into one item. Therefore, the bias of $HSIC_b$ in Equation (17) can be written as:

$$bias(HSIC_b) = \mathbb{E}(HSIC_b - HSIC) = \frac{1}{m}(\mathbb{E}_{xy}kl$$
$$- 2\mathbb{E}_{xyy'}kl - 2\mathbb{E}_{xx'y}kl + \mathbb{E}_{xy'y"}kl + \mathbb{E}_{xx'y"}kl$$
$$+ \mathbb{E}_{xx'}k\mathbb{E}_{yy'}l) + O(m^{-2}) = E(HSIC_0) + O(m^{-2}). \quad (20)$$

Therefore, the deviation of PHSIC (Z) in independent cases as:

$$bias(PHSIC(Z)) = \mathbb{E}(PHSIC - HSIC) = O(m^{-2}). \quad (21)$$

□

Corollary 2. *The deviation of $PHSIC(Z, \mathcal{F}, \mathcal{G})$ in dependent cases is $O(m^{-1})$.*

Proof. In dependence cases, the deviation of PHSIC is shown in equation (19), we can see that

$$bias(PHSIC(Z)) = \mathbb{E}(PHSIC - HSIC) = O(m^{-1}). \quad (22)$$

□

3.1 Advantages of PHSIC Compared to HSIC

In this section, we mainly analyze the advantages of PHSIC compared to $HSIC_b$ in independent situations.

PHSIC has zero baseline. In the independent case, we can see that the mean of PHSIC is as follows:

$$\mathbb{E}(PHSIC(Z)) = \mathbb{E}((HSIC_b(Z) - \mathbb{E}(HSIC_0(Z)) = 0. \quad (23)$$

And the average value of $HSIC_b$ under independent conditions is shown in equation (7), its mean is not 0. Therefore, PHSIC is more interpretable.

PHSIC is closer to unbiased in sample size. From Theorem 1, we deduce that $HSIC_b$ has a deviation order of $O(m^{-1})$ in the independent case, while PHSIC has a deviation order of $O(m^{-2})$ in the independent case (as shown in equation (21)). Therefore, the latter has a smaller error order than the former and is closer to unbiased.

PHSIC may closer to unbiased in dimensions. From equation (21), it can be seen that the deviation term of $O(m^{-1})$ in $HSIC_b$ comes from $E(HSIC_0)$. We observe the dimensional variation of this term from Monte Carlo perspective. Assuming that $X^{m\times d}$, $Y^{m\times d}$ is independent and uniform multidimensional random variables, d represents dimension. $X = [x(1), x(2)..., x(d)]$. The values of each dimension of X are the same, but the positions are different. $Y = X(\pi)$, π indicates disrupting the order of column X. In this way, X and Y can remain independent, and the values of the number of columns are the same, but the positions are different. for $||u_{x^d}||^2 = (m)_2^{-1}\sum_{(i,j)\in i_2^m} k_{i,j}$, when m is fixed, the individual term k_{ij} decreases as the dimension d increases, since the value of $||u_{x^d}||^2$ reduce. Therefore, for $\mathbb{E}(HSIC_0(Z_n|X,Y)) = \frac{1}{m}(1 + ||u_x||^2||u_y||^2 - ||u_x||^2 - ||u_y||^2)$, $(m)_2^{-1}\sum_{(i,j)\in i_2^m} k_{i,j} \in (0, 1]$. We think of it as a function of $||u_x||^2$ (because the values of $||u_x||^2$ and $||u_y||^2$ are the same). Let $||u_x||^2 = a$, $\mathbb{E}(HSIC_0(Z)) = b$ is a function of a, $b = 1 + a^2 - 2a$, derive from it, $b' = 2(a-1)$, $a \in (0, 1]$, b is a monotonically decreasing function, When a decreases, b is monotonically increasing. Therefore, the $\mathbb{E}(HSIC_0(Z)$ is monotonically increasing as the variable dimension increases. PHSIC subtracts $E(HSIC_0)$ from $HSIC_b$, making the dimension closer to unbiased.

According to the above analysis, from the perspective of random consistency, PHSIC has a smaller deviation compared to HSIC and has stronger properties.

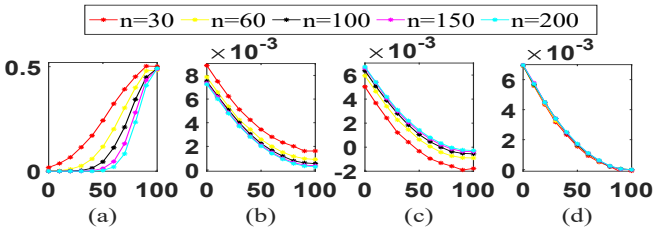


Figure 2: Scores of (a) pHSIC, (b) $HSIC_b$, (c) SHSIC, and (d) PHSIC under different samples and noise conditions. The horizontal axis: Noise level , and the vertical axis: score.

3.2 Toy Example

We establish toy examples to further understand the advantages of PHSIC. We compare PHSIC with pHSIC (p-value of $HSIC_b$) and a popular depolarization framework SHSIC. Equation (24) is the calculation formula for SHSIC, a is constant with the value of 0.5.

$$SHSIC(Z) = HSIC_b - \mathbb{E}(HSIC_0) - \sqrt{\frac{1-a}{a}Var(HSIC_0)}. \quad (24)$$

Figure 2 shows the scores of pHSIC, $HSIC_b$, SHSIC and PHSIC under different noise levels in the nonlinear function $y = 4(x-5)^2$ in the case of binary univariate and random kernel parameters. It can be found from the four figures that compared with the other three indicators, PHSIC not only has zero baseline, but also converges quickly with a small sample. We find that p-values are only effective under global distribution, and in some smaller samples, p-values will be higher under correlated conditions, which leads to a type II error and judges some correlated values as independent. From Figure 3, we observe the sample selection fairness of PHSIC in a more intuitive way. We generated 5 groups of samples with different quantities to simulate the correlation between variables. The estimated values of S_n for different samples are calculated to calculate their selection probabilities. We select samples according to $HSIC_b$, PHSIC and SHSIC, and observe the fairness of these estimators in selecting samples of different quantities. For Figure 3 (a) and (b), we set Sn=[120 140 160 180 200] and simulated the correlations of x and y under correlated and independent conditions. Observing the selection probabilities of SHSIC, PHSIC, and $HSIC_b$ for these five groups of samples, we can clearly see that PHSIC has more selection fairness.

In order to verify the unbiased nature of dimensions, we conduct experiments on independent uniform distribution simulation data with a sample size of 300 and a value range of 11-12. We increase the dimensions of x and y from 1 to 5 and compare the scores generated by different dimensions. From Figure 3 (c), it can be seen that PHSIC is more unbiased compared to $HSIC_b$ and SHSIC, and its selection probability in different dimensions is closer to 0.2, $HSIC_b$ and SHSIC tend to be more high-dimensional, which proves that PHSIC has more fair selectivity in different dimensions.

To further verify the fairness of PHSIC, we evaluate it under different samples and functional relationships. As shown by the red line in Figure 4 (a), we use $HSIC_b$ to measure the

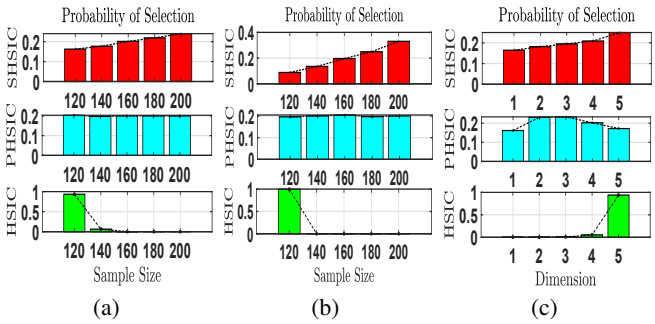


Figure 3: The selection fairness of SHSIC, PHSIC, and $HSIC_b$ in different dimensions and samples.

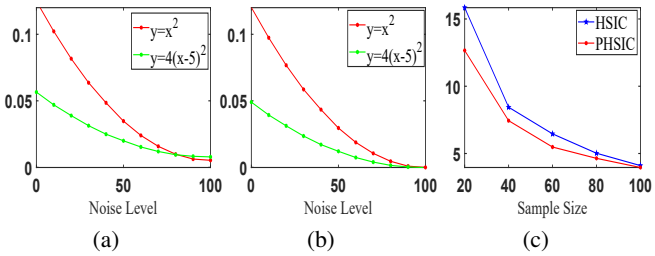


Figure 4: Figures (a) and (b) represent the scores of $HSIC_b$ and PHSIC under different cost and functional relationships, respectively. Figure (c) shows the probability that the values of PHSIC and $HSIC_b$ in independent cases are greater than those in correlated cases under the same sample size.

correlation metric scores of variables x and y simulated by the function $y = x^2$ under different noise levels for 60 samples. The green line represents the correlation metric scores of variables x and y simulated by the function $y = 4(x-5)^2$ under different noise levels when $HSIC_b$ measures 40 samples. Figure 4 (b) using PHSIC measurement as a control. From the functional relationship, it can be seen that the scores of the variables x and y simulated by $y = x^2$ are always higher than those simulated by $y = 4(x-5)^2$ in non independent cases, and the values should be the same in independent cases. From Figure 4(a), it can be seen that the independent values of $HSIC_b$ measures are higher than those in some non independent cases, leading to false associations. From Figure 4 (b), it can be seen that PHSIC effectively solves this situation.

In the above toy example, we mainly demonstrate the fairness of PHSIC in different and identical associations under different sample sizes. Next, we further demonstrate its robustness in the same sample and different association relationships. We use $HSIC_b$ and PHSIC to measure the scores of variables x and y in independent and non independent cases under the same sample size. We simulate 10000 times and observe the probability that the scores of two indicators were higher in independent cases than in non independent cases. From Figure 4 (c), we can see that $HSIC_b$ are more prone to misselection in the case of limited samples, further demonstrating the good performance of PHSIC.

DATASET	$HSIC_{x\to y}$	$HSIC_{x\leftarrow y}$	DECISION	$PHSIC_{x\to y}$	$PHSIC_{x\leftarrow y}$	DECISION	GROUND TRUTH
1	0.0269	0.0153	$x \leftarrow y$	0.0037	0.0104	$x \to y$	$x \to y$
2	0.0231	0.0195	$x \leftarrow y$	0.0006	0.0184	$x \to y$	$x \to y$
3	0.0225	0.0121	$x \leftarrow y$	0.0000	0.0601	$x \to y$	$x \to y$
4	0.0253	0.0164	$x \leftarrow y$	0.0010	0.0064	$x \to y$	$x \to y$
5	0.0751	0.0703	$x \leftarrow y$	0.0278	0.0522	$x \to y$	$x \to y$
6	0.1660	0.0829	$x \leftarrow y$	0.0299	0.0356	$x \to y$	$x \to y$
7	0.0238	0.0569	$x \to y$	0.0269	0.0159	$x \leftarrow y$	$x \leftarrow y$
8	0.0229	0.0589	$x \to y$	0.0291	0.0122	$x \leftarrow y$	$x \leftarrow y$

Table 1: Experimental results on ANM-MM Model.

DATASET	$HSIC_{x\to y}$	$HSIC_{x\leftarrow y}$	DECISION	$PHSIC_{x\to y}$	$PHSIC_{x\leftarrow y}$	DECISION	GROUND TRUTH
1	0.0800	0.0788	$x \leftarrow y$	0.0100	0.0156	$x \to y$	$x \to y$
2	0.0691	0.0829	$x \to y$	0.0348	0.0221	$x \leftarrow y$	$x \leftarrow y$
3	0.1628	0.1029	$x \leftarrow y$	0.0409	0.0496	$x \to y$	$x \to y$
4	0.0697	0.0660	$x \leftarrow y$	0.0231	0.0379	$x \to y$	$x \to y$
5	0.0569	0.0536	$x \leftarrow y$	0.0095	0.0384	$x \to y$	$x \to y$
6	0.0384	0.1106	$x \to y$	0.0267	0.0129	$x \leftarrow y$	$x \leftarrow y$
7	0.1191	0.0570	$x \leftarrow y$	0.0163	0.0067	$x \leftarrow y$	$x \leftarrow y$
8	0.1143	0.1074	$x \leftarrow y$	0.0288	0.0341	$x \to y$	$x \leftarrow y$

Table 2: Experimental results on KIKO Model.

DATASET	HSIC			PHSIC			DATASET	HSIC		PHSIC	
	6	7	8	6	7	8		2	3	2	3
$y = e^x$	43%	82.25%	70.25%	**44.5%**	**83.75%**	**71.25%**	ozon	**88.5%**	74.6%	**88.5%**	**79%**
$y = x$	10.16%	56.28%	45.6%	**18.16%**	**59%**	**51.6%**	$y = x^2$	**100%**	**100%**	**100%**	**100%**
$y = sin(x)$	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	mpg-acc	**81%**	64.3%	80%	**69.6%**
$concrete$	**83%**	78%	79%	**83%**	**80.7%**	**79.5%**	mpg-mpg	77%	84%	**81.5%**	**86.6%**

Table 3: Experimental results on HANM Model.

4 PHSIC-Based Causal Inference Models

In this section, we first introduce several typical causal inference models, then describe the data and experimental details we used for evaluation, and finally present the experimental results.

4.1 Introduce Typical Causal Inference Models

In this paper, we mainly focus on four classic causal inference models: ANM [Hoyer *et al.*, 2008], ANM-MM[Hu *et al.*, 2018], KIKO [Assaad *et al.*, 2019], HANM [Zhao *et al.*, 2023].

- ANM-MM model. The additive noise model is extended to a mixed model consisting of a finite number of ANMs, and its causal identifiability conditions are given.
- KIKO model. In the context of ANM model, KIKO model uses one regression variable instead of two to accelerate causal inference.
- HANM model. Identify many to one causal relationships and use asymmetric forward and backward models HANM to identify causal direction.
- ANM model. Inferring causal directions through the asymmetry of forward and backward models using nonlinear functions.

All these models are based on HSIC to measure the independence of causes and residuals, which is a key part of these methods. Therefore, next we will use PHSIC to further improve the performance of these models.

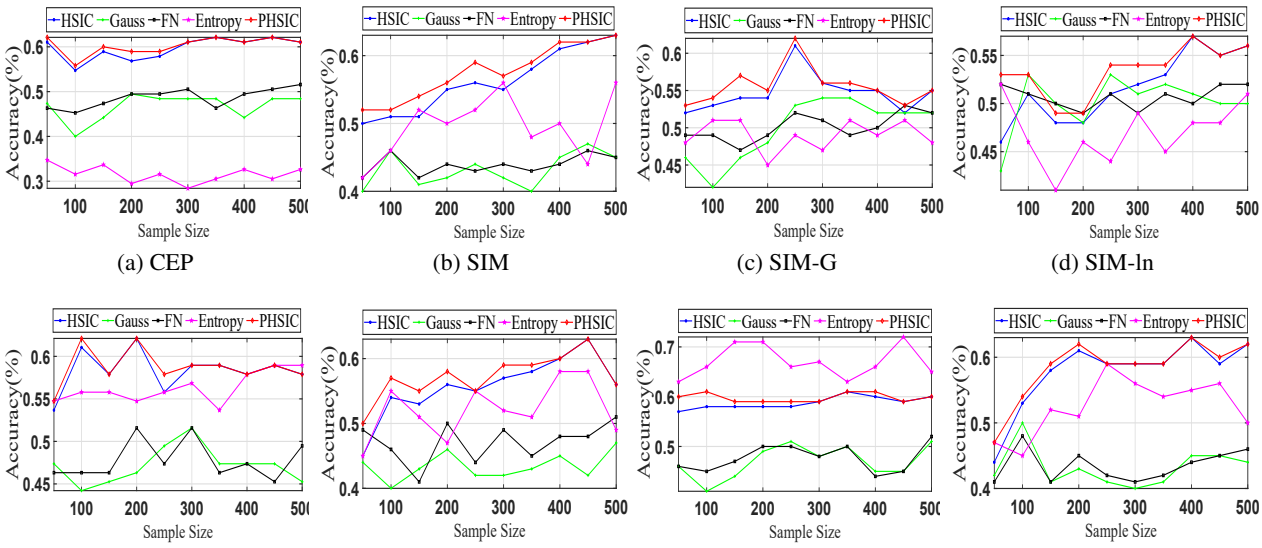
4.2 Experimental Setup

In this section, we introduce the specific experimental details of the four models in the appeal:

ANM-MM model. we use eight real datasets: (Rings, Viscera), (mpg, Weight) and (age, weight), etc. In order to highlight the advantages of our PHSIC, we conducted experiments on 50 samples with the same experimental parameters as the original ANM-MM.

KIKO model. we use eight real datasets:(Temperature, Co2 flux),(GNI, life expectancy) and (NEP, PPFDdif), etc. And we conducted experiments on 30 samples with the same experimental parameters as the original KIKO model.

HANM model. we use four simulated data and four real data. For the first four data ($y = e^x$, $y = x$, $y = sin(x)$, $concrete$), we conducted experiments on 6, 7, and 8 causes respectively, and for the last four data (ozon, $y = x^2$, mpg-acc, mpg-mpg), we use 2 and 3 causes respectively. Exponential data($y = e^x$) and $concrete$ data were tested on 50 samples, while others were tested on 30 samples.



(a) CEP (b) SIM (c) SIM-G (d) SIM-ln

(e) CEP(10% Gaussian noise) (f) CEP(10% uniform noise) (g) SIM(10% Gaussian noise) (h) SIM (10% uniform noise)

Figure 5: Experimental results on ANM Model.

ANM model. We use one real dataset(CEP) and three simulated datasets(SIM, SIM-G, SIM-IN). We compare the methods based on HSIC score (ANM-HSIC), Entropy score (ANM-Entropy), Gaussian score (ANM-Gauss), and empirical Bayesian score (ANM-FN) with our proposed PHSIC score (ANM-PHSIC). The introduction of the above comparison methods can be found in [Mooij *et al.*, 2016].

A more detailed introduction to the above dataset and experimental parameters will be provided in the appendix. For HANM and ANM models, we use accuracy indicators for evaluation, and the calculation method is as follows:

$$accuracy = \frac{\sum_{m=1}^{M} w_m \delta_{\hat{d}_m, d_m}}{\sum_{m=1}^{M} w_m}, \tag{25}$$

where w_m is the weight of each pair of data, because there is a strong correlation between variables in the real dataset, CEP dataset decouples each pair of variables by assigning different weights. All pairs from the same dataset are weighted equally and sum to 1. d_m and $\hat{d}_m$ are the real direction and predicted direction of the mth pair, respectively, if the two directions are the same, it is recorded as 1, otherwise it is 0.

4.3 Experimental Result

Table 1 shows the experimental results of the HSIC based ANM-MM model and the PHSIC based ANM-MM on eight datasets. Compared with GROUND TRUTE, it can be found that the ANM-MM model based on PHSIC can be correctly recognized on all eihgt datasets, while the ANM-MM model based on HSIC cannot be correctly recognized due to its high bias in small samples. Table 2 shows the experimental results of the HSIC based KIKO model and the PHSIC based KIKO model on eight datasets. Compared with GROUND TRUTE, it can be observed that the KIKO model based on PHSIC can correctly recognize on most datasets. Table 3 shows the experimental results of the HSIC based HANM model and the PHSIC based HANM mode on eight datasets. It can be found that the HANM model based on PHSIC has significant advantages in identifying many to one causal relationships on small samples.

Figure 5 shows the experimental results on the ANM model. Figures 5 (a), (b), (c), and (d) show the experimental results of different methods on different sample sizes on the datasets CEP, SIM, SIM-G and SIM-ln, respectively. The accuracy of ANM-PHSIC in small samples is mostly higher than other indicators. As the number of samples increases, the accuracy of ANM-HSIC is consistent with that of ANM-PHSIC. To verify the robustness of ANM-PHSIC under different noises, we add 10% gaussian noise and independent uniform noise to CEP and SIM data respectively. As shown in Figure 5 (e), (f), (g), (h), ANM-PHSIC also has strong robustness compared to other methods in most cases. In summary, ANM-PHSIC is more robust than other methods in cases of sparse samples and noise data.

5 Conclusion

This paper proposes a pure HSIC metric (PHSIC) to against random consistency, proving that PHSIC has tighter bounds than HSIC, and verifying its fairness and stability through theoretical analysis and simulation experiments. In addition, we apply PHSIC to multiple causal inference tasks, improving the robustness of these models in small sample and noisy scenarios. Furthermore, PHSIC can be applied to machine learning, deep learning and other fields. This is also the direction we will further apply in the future.

In conclude, it also gave us an inspiration. Unlike the traditional definition of independence when the sample tends to infinity, independence should be data-driven. From the perspective of random consistency, we can redefine the independence between random variables to construct robust and fair correlation indicators.

Acknowledgments

This work was supported by National Key Research and Development Program of China (No. 2021ZD0112400), the National Natural Science Foundation of China (Nos.62136005), the Science and Technology Major Project of Shanxi (No.202201020101006), the Key $R\&D$ Program of Shanxi Province, China (Grant no.202202020101004), the Natural Science Foundation of Shanxi Province, China (Grant nos.20210302123455).

References

[Assaad *et al.*, 2019] Charles Karim Assaad, Emilie Devijver, Eric Gaussier, and Ali Ait-Bachir. Scaling causal inference in additive noise models. In *The 2019 ACM SIGKDD Workshop on Causal Discovery*, pages 22–33. PMLR, 2019.

[Gretton *et al.*, 2005] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[Gretton *et al.*, 2007] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

[Hotelling, 1992] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pages 162–190. Springer, 1992.

[Hoyer *et al.*, 2008] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

[Hu *et al.*, 2018] Shoubo Hu, Zhitang Chen, Vahid Partovi Nia, Laiwan Chan, and Yanhui Geng. Causal inference and mechanism clustering of a mixture of additive noise models. *Advances in neural information processing systems*, 31, 2018.

[Jaworski *et al.*, 2010] Piotr Jaworski, Fabrizio Durante, Wolfgang Karl Hardle, and Tomasz Rychlik. *Copula theory and its applications*, volume 198. Springer, 2010.

[Li *et al.*, 2021] Guohe Li, Yong Li, Yifeng Zheng, Ying Li, Yunfeng Hong, and Xiaoming Zhou. A novel feature selection approach with pareto optimality for multi-label data. *Applied Intelligence*, pages 1–18, 2021.

[Liaghat and Mansoori, 2019] Samaneh Liaghat and Eghbal G Mansoori. Filter-based unsupervised feature selection using hilbert–schmidt independence criterion. *International journal of machine learning and cybernetics*, 10:2313–2328, 2019.

[Ma *et al.*, 2020] Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5085–5092, 2020.

[Mooij *et al.*, 2016] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

[Niu *et al.*, 2010] Donglin Niu, Jennifer G Dy, and Michael I Jordan. Multiple non-redundant spectral clustering views. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 831–838, 2010.

[Ogarrio *et al.*, 2016] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Conference on probabilistic graphical models*, pages 368–379. PMLR, 2016.

[Puth *et al.*, 2015] Marie-Therese Puth, Markus Neuhäuser, and Graeme D Ruxton. Effective use of spearman's and kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, 102:77–84, 2015.

[Romano *et al.*, 2016] Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. A framework to adjust dependency measure estimates for chance. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 423–431. SIAM, 2016.

[Schweizer and Wolff, 1981] Berthold Schweizer and Edward F Wolff. On nonparametric measures of dependence for random variables. *The annals of statistics*, 9(4):879–885, 1981.

[Shimizu *et al.*, 2006] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[Song *et al.*, 2007a] Le Song, Alex Smola, Arthur Gretton, and Karsten M Borgwardt. A dependence maximization view of clustering. In *Proceedings of the 24th international conference on Machine learning*, pages 815–822, 2007.

[Song *et al.*, 2007b] Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 823–830, 2007.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[Vinh *et al.*, 2009] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080, 2009.

[Wang *et al.*, 2022] Jieting Wang, Yuhua Qian, Feijiang Li, Jiye Liang, and Qingfu Zhang. Generalization performance of pure accuracy and its application in selective ensemble learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1798–1816, 2022.

[Yamanishi *et al.*, 2004] Yoshihiro Yamanishi, Jean-Philippe Vert, and Minoru Kanehisa. Heterogeneous data comparison and gene selection with kernel canonical correlation analysis. *Kernel methods in computational biology*, pages 209–229, 2004.

[Zhang and Hyvarinen, 2012] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

[Zhao *et al.*, 2023] Boxiang Zhao, Shuliang Wang, Lianhua Chi, Chuanfeng Zhao, Hanning Yuan, Qi Li, Xiaojia Liu, Jing Geng, and Ye Yuan. Hanm: Hierarchical additive noise model for many-to-one causality discovery. *IEEE Transactions on Knowledge and Data Engineering*, 2023.