Multi-Level Wavelet Mapping Correlation for Statistical Dependence Measurement: Methodology and Performance

Yixin Ren[1], Hao Zhang[1], Yewei Xia[1], Jihong Guan[2], Shuigeng Zhou[1*]

[1]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China
[2]Department of Computer Science & Technology, Tongji University, China
{yxren21, ywxia21}@m.fudan.edu.cn, {haoz15, sgzhou}@fudan.edu.cn, jhguan@tongji.edu.cn

Abstract

We propose a new criterion for measuring dependence between two real variables, namely, Multi-level Wavelet Mapping Correlation (MWMC). MWMC can capture the nonlinear dependencies between variables by measuring their correlation under different levels of wavelet mappings. We show that the empirical estimate of MWMC converges exponentially to its population quantity. To support independence test better with MWMC, we further design a permutation test based on MWMC and prove that our test can not only control the type I error rate (the rate of false positives) well but also ensure that the type II error rate (the rate of false negatives) is upper bounded by O(1/n) (n is the sample size) with finite permutations. By extensive experiments on (conditional) independence tests and causal discovery, we show that our method outperforms existing independence test methods.

Introduction

Detecting dependence between two variables is a fundamental problem in the field of statistics, with applications in a variety of areas such as statistical inference (Casella and Berger 2021), independent principal component analysis (Comon 1994), and feature selection (Fukumizu, Bach, and Jordan 2004). When two random variables are categorical, a class of non-parametric methods such as Pearson's chi-square test (Plackett 1983), are used to determine whether they are independent. However, when the density functions of two random variables are both continuous, determining the independence between them is considered to be a challenging task (Hoeffding 1948).

Pearson's correlation coefficient (Benesty et al. 2009) is widely used as a criterion for testing the independence of continuous variables. However, since it can only measure linear correlations between variables, there is no guarantee that all dependencies can be measured when the distributions of variables are non-Gaussian. In order to detect more complex nonlinear dependencies among random variables, a class of kernel-based (Muller et al. 2001) independence criteria were proposed based on the framework established by Rényi (Rényi 1959). These criteria are mainly derived from the cross-covariance operators in the reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan 2011). The first proposed RKHS measure is the kernel canonical correlation (KCC) (Bach and Jordan 2002), which mainly uses the maximum singular value to measure dependency. Later, the constrained covariance (COCO) (Gretton et al. 2005b) criterion without regularization was proposed. Some other empirical kernel quantities such as kernel mutual information (KMI) (Gretton, Herbrich, and Smola 2003) and kernel generalized variance (KGV) (Bach and Jordan 2002) use Parson's window to estimate mutual information to measure dependence. One of the most widely used kernel-based dependence measures, the Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al. 2005a), uses the squared *Hilbert-Schmidt norm* to detect dependence. The HSIC-based independence test outperforms other RKHS measures in experimental performance with adequate feature mapping. However, this kernel-based method has at least quadratic computational complexity and has to choose the kernel bandwidth. Some pointwise distance based methods (Székely and Rizzo 2009; Heller, Heller, and Gorfine 2013; Lyons 2013) and copula based methods (Schweizer and Wolff 1981; Zhang 2019) are also widely used. A representative copula based method, the randomized dependence coefficient (Lopez-Paz, Hennig, and Schölkopf 2013) (RDC), uses nonlinear random projections to maximize the canonical correlation of the copula transformations corresponding to the random variables. Recently, Chatterjee (Chatterjee 2021) proposed a correlation coefficient capable of measuring the strength of the nonlinear correlation between two variables.

On the other hand, deep neural networks have also been recently proposed to test independence. He et al. (He et al. 2021) developed a neural network-based independence test method, which was used in their proposed continuous optimization algorithm DARING. For simplicity, we treat DARING and the proposed independence test method equally in the following. DARING optimizes the parameterized function $f(x)$ (x is a random variable) to maximize the correlation coefficient between $f(x)$ and the random variable y by using deep neural network. However, zero correlation between $f(x)$ and y is theoretically insufficient to support independence according to the previous work (Daudin 1980). In contrast to kernel-based independence tests, DARING is not stable/robust, and is not easy to be applied to new scenarios, due to a bunch of hyperparameters.

In this work, we propose a novel computational dependency criterion called **M**ulti-level **W**avelet **M**apping **C**orrelation (MWMC). By decomposing the L^2 space into mutually orthogonal wavelet spaces, we convert nonlinear dependencies into correlations under wavelet mappings. Thanks to the sufficient scaling and translation properties of wavelet, we circumvent the problem of bandwidth selection. Concretely, our contributions are summarized as follows:

- We propose a new dependency criterion MWMC that can capture the nonlinear dependency between two real variables by measuring their correlation under different levels of wavelet mappings.
- We show that the empirical estimate of MWMC converges to its population quantity at a rate of $O(n^{-1/2})$ where n is the sample size, thus ensuring that MWMC is a practical criterion for independence test.
- We design a permutation test based on MWMC for testing independence, and prove that its type I error rate can be well controlled meanwhile the type II error rate is upper bounded by $O(n^{-1})$ with finite permutations.
- We conduct extensive experiments, and the results show that our method outperforms the state of the art methods. In particular, our method can handle better the situation where two variables are uncorrelated but not independent, and the signal-to-noise ratio is small. This is one of the "hardest" cases in independence test, where most existing methods are prone to type II errors.

The rest of this paper is organized as follows: we first introduce the basic concepts of hypothesis testing and wavelet analysis in the preliminaries. Then we define MWMC and give its corresponding estimator, and establish the convergence relation between them. After that, we propose the permutation test using MWMC and present our main theoretical results (including the bound on type II error). The experiments on synthetic and real data are given in the performance evaluation section. Finally, we conclude the paper.

Preliminaries

Hypothesis Testing

We consider two random variables x and y in a one-dimensional probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$[1] and assume that the observed data consist of n *i.i.d.* pairs (x_i, y_i). The problem of testing independence between x and y can be written in the form of a hypothesis testing:

$$\mathcal{H}_0 : x \perp\!\!\!\perp y \quad \text{versus} \quad \mathcal{H}_1 : x \not\perp\!\!\!\perp y. \tag{1}$$

Independence hypothesis testing generally consists of the following steps. First, state the relevant statistic T and calculate the observed value of T from the observational data. Then, give a user-selected significance level α (typically taken as 0.05), which indicates the lowest limit of the probability threshold for rejecting $\mathcal{H}_0$. After that, obtain the p-value, which is the probability that the sampling of T under $\mathcal{H}_0$ is at least as extreme as the observed value. Finally, the null hypothesis $\mathcal{H}_0$ is rejected if the p-value is not greater than α. There are two types of errors may occur during hypothesis testing. Type I error means the rejection of $\mathcal{H}_0$ when it is true, and Type II error indicates when $\mathcal{H}_0$ is wrong but not rejected. A well-performed independence test requires that Type I error rate is upper bounded by α meanwhile Type II error is minimized (Zhang et al. 2011).

[1]As in the literature, here $\mathcal{B}(\mathbb{R})$ stands for the Borel σ-algebra on $\mathbb{R}$ and λ stands for the Lebegue probability measure.

Wavelet Analysis

Wavelet transform (Pavlov et al. 2012) has a wide range of applications in signal processing. One of its most exciting properties is the ability to decompose arbitrary square integrable (equivalent to finite energy) functions into wavelet combinations of different levels, which are generated by scaling and shifting of the mother wavelet function (Torrence and Compo 1998). In this paper, we mainly consider a class of compactly supported scaling functions, for instance, B-spline as a mother wavelet.

Take the linear B-spline wavelet ϕ as an example, it is defined as follows[2]:

$$\phi(x) = \begin{cases} 1 - |x| & \text{for} \quad -1 \le x \le 1 \\ 0 & \text{for} \quad \text{elsewhere} \end{cases}. \tag{2}$$

Then, we define the scaling space $V_0 := \{\sum_{k\in\mathbb{Z}} a_k \phi(x-k) \,\big|\, a_k \in \mathbb{R}\}$ and the wavelet space at level j, $W_j := \{\sum_{k\in\mathbb{Z}} b_k \psi(2^j x - k) \,\big|\, b_k \in \mathbb{R}, j \ge 0\}$, where $\psi(x) = \phi(2x+1) - \phi(2x-1)$ indicates the orthogonal complement of $\phi(x)$. For simplicity, we denote $\phi(x-m)$ as $\psi_0(x-m)$ and $\psi(2^j x - m)$ as $\psi_{j+1}(x-m)$ in the rest of the paper. And for convenience of understanding, we restate one of the key properties of wavelets in Theorem 1 by following (Daubechies 1992, Chap. 5):

Theorem 1. (Wavelet Decomposition) *Let $\mathcal{L}^2(R)$ be the space of square integrable functions, V_0 be the scaling space, W_j be the wavelet space at level j, then $\mathcal{L}^2(R) = V_0 \oplus W_0 \oplus W_1 \oplus \cdots$.*

Multi-level Wavelet Mapping Correlation

In this section, we first introduce MWMC as a new dependence measure. Then, we give an estimator for MWMC with finite samples and show that this estimate converges to MWMC quickly.

MWMC Criterion

Definition 1. (Wavelet Mapping Covariance (WMCov)). *Let x, y be two continuous random variables (RVs) with joint probability density function $P_{x,y}$. Then, for any levels $i, j \in \mathbb{N}$ and shiftings $m, k \in \mathbb{Z}$, we define the covariance between the wavelet mappings $\psi_i(x-m)$ and $\psi_j(y-k)$ as*

$$\begin{aligned} Cov^{\psi}_{ij,mk}(x,y) :=& \mathbf{E}_{x,y}[\psi_i(x-m)\psi_j(y-k)] \\ &- \mathbf{E}_x[\psi_i(x-m)]\mathbf{E}_y[\psi_j(y-k)]. \end{aligned} \tag{3}$$

As stated in Theorem 2 with a proof given in Appendix, the dependence between random variables can be captured

[2]Compared to the general convention, we make slight changes to facilitate the rest of the derivation in this paper. And, we also refer to this wavelet ϕ as a scaling function.

by covariances of wavelet mappings of different scales and shiftings.

Theorem 2. (Independence). *The RVs x and y are independent if and only if* $Cov^{\psi}_{ij,mk}(x,y) = 0$ *for all levels* $i, j \in \mathbb{N}$ *and shiftings* $m, k \in \mathbb{Z}$.

Due to the large difference in the norm of wavelet functions of different levels, we need a normalized coefficient. For levels $i, j \in \mathbb{N}$, we define **W**avelet **M**apping **Cor**relation coefficient (WMCor) ρ^{ψ}_{ij} as

$$\rho^{\psi}_{ij} = \sup_{m,k\in\mathbb{Z}} \rho^{\psi}_{ij,mk} = \sup_{m,k\in\mathbb{Z}} \left| \frac{\mathrm{Cov}^{\psi}_{ij,mk}(x,y)}{\sqrt{\mathrm{Var}^{\psi}_{i,m}(x)+\kappa}\sqrt{\mathrm{Var}^{\psi}_{j,k}(y)+\kappa}} \right|, \tag{4}$$

where $\mathrm{Var}^{\psi}_{i,m}(x) := \mathbf{E}_x[\psi_i^2(x-m)] - \mathbf{E}_x[\psi_i(x-m)]^2$ and κ is a small positive constant to prevent the denominator of ρ^{ψ}_{ij} from being 0 (which may occur due to the compact support property of the wavelet). WMCor inherits the independence characterization property of WMCov.

By defining $\rho^{\psi} := \sup_{i,j\in\mathbb{N}} \rho^{\psi}_{ij}$, then we have that the random variables x and y are independent if and only if $\rho^{\psi} = 0$. However, it is impossible to calculate ρ^{ψ} by traversing wavelet functions of a countable number levels in practice. We take the first l levels of ρ^{ψ} as its approximation and obtain **M**ulti-level **W**avelet **M**apping **C**orrelation coefficient (MWMC) $\rho^{\psi}_{l} := \sup_{0\le i\le l, 0\le j\le l} \rho^{\psi}_{ij}$.

We will prove that as the level l increases, MWMC ρ^{ψ}_{l} approximates ρ^{ψ} under the assumption of smoothness of the probability density function.

Follow the notion of lipschitz density introduced in (Qi 2020), we have the following definition:

Definition 2 (L-lipschitz RV Pair)**.** *Let* x, y *be the continuous random variables (RVs) with joint probability density function* $P_{x,y}$. *Then,* (x, y) *is defined as a L-lipschitz RV Pair if there exists a finite constant L, such that* $P_{x,y}$ *satisfies* $|P_{x,y}(x_1,y_1) - P_{x,y}(x_2,y_2)| \le L\|(x_1,y_1)-(x_2,y_2)\|_2$ *for almost every points* $(x_1,y_1),(x_2,y_2) \in \mathbb{R}\times\mathbb{R}$.

The L-lipschitz condition in Def. 2 limits the smoothness of the distribution. And since in practice we usually normalize the distribution in preprocessing, we assume that the random variables discussed in the rest of the paper take values in the interval $[-1, 1]$. This assumption has no significant impact in practice, see Appendix for more detailed explanation. Then, the difference between ρ^{ψ} and ρ^{ψ}_{l} is determined by l, as presented in Lemma 1 with a proof given in Appendix.

Lemma 1. *Let* (x, y) *be a L-lipschitz RV pair and each variable takes values in the interval* $[-1, 1]$. *Then, for all* $\epsilon > 0$, *exist* $l^* = \lceil \log_2(\frac{1+12L}{\epsilon\kappa}) \rceil + 1$, *such that for all* $l \ge l^*$, $|\rho^{\psi}_{l} - \rho^{\psi}| \le \epsilon$.

Roughly speaking, when ϵ is small, the required level l is asymptotically $O(\log(1/\epsilon))$.

Estimator of MWMC

Here, we establish the relationship between MWMC and its estimator.

Definition 3. (Empirical estimate of MWMC, or Empirical MWMC in short). *Let* $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ *be the set of n i.i.d. pairs* (x_i, y_i) *drawn from the joint probability density function* $P_{x,y}$. *Then, for finite level* $l \in \mathbb{N}$, *an estimator of MWMC is given by*

$$\begin{aligned}\widehat{\rho}^{\psi}_{l} &= \sup_{0\le i,j\le l} \widehat{\rho}^{\psi}_{ij} = \sup_{0\le i,j\le l} \sup_{m,k\in\mathbb{Z}} \widehat{\rho}^{\psi}_{ij,mk} \\ &= \sup_{0\le i,j\le l} \sup_{m,k\in\mathbb{Z}} \left| \frac{\widehat{Cov}^{\psi}_{ij,mk}(x,y)}{\sqrt{\widehat{Var}^{\psi}_{i,m}(x)+\kappa}\sqrt{\widehat{Var}^{\psi}_{j,k}(y)+\kappa}} \right|, \end{aligned} \tag{5}$$

where $\widehat{\mathrm{Cov}}^{\psi}_{ij,mk}(x,y) := \frac{1}{n}\sum_p[\psi_i(x_p-m)\psi_j(y_p-k)] - \frac{1}{n(n-1)}\sum_{p\ne q}[\psi_i(x_p-m)\psi_j(y_q-k)]$ is an unbiased estimate of $\mathrm{Cov}^{\psi}_{ij,mk}(x,y)$ and $\widehat{\mathrm{Var}}^{\psi}_{i,m}(x) := \frac{1}{2n(n-1)}\sum_{p\ne q}[\psi_i(x_p-m) - \psi_i(x_q-m)]^2$ is the corresponding unbiased estimate of $\mathrm{Var}^{\psi}_{i,m}(x)$. We have the following theorem to establish the connection between $\widehat{\rho}^{\psi}_{l}$ and ρ^{ψ}_{l}, with a proof given in Appendix.

Theorem 3. (Bound on Empirical MWMC). *Assume* $\kappa < 0.1$, *then for* $n > 1$, *finite level* $l \in \mathbb{N}$ *and all* $\delta > 0$, *with probability at least* $1-\delta$, *the following bound holds*

$$|\widehat{\rho}^{\psi}_{l} - \rho^{\psi}_{l}| \le \sqrt{\frac{\log(4/\delta)+2l+3}{\kappa^4}}\frac{12}{\sqrt{n}}. \tag{6}$$

Theorem 3 indicates that the difference between the Empirical MWMC and its population value is very small under the condition of sufficient samples. Combining the results in Lemma 1 and Theorem 3, it is easy to derive the following corollary:

Corollary 1. *Let* (x, y) *be a L-lipschitz RV pair and each variable takes values in the interval* $[-1, 1]$. *Assume* $\kappa < 0.1$, *then for all* $\epsilon > 0$, *there exist* $l^* = \lceil \log_2(\frac{1+12L}{\epsilon\kappa}) \rceil + 1$, $n^* = \lceil \frac{144[\log(4/\delta)+2l^*+3]}{\epsilon^2\kappa^4} \rceil$, *such that for all* $l \ge l^*, n \ge n^*$, $|\widehat{\rho}^{\psi}_{l} - \rho^{\psi}| \le 2\epsilon$ *with probability at least* $1-\delta$.

Therefore, when the sample size n and the level l are large enough, the difference between Empirical MWMC $\widehat{\rho}^{\psi}_{l}$ that we use in practice and the ground truth value ρ^{ψ} can be controlled within an acceptable accuracy range with high probability.

In order to use $\widehat{\rho}^{\psi}_{l}$ for independence test, we need to estimate p-value. A common resampling method is to perform a permutation test, which we discuss in the next section.

Permutation Tests for MWMC

In this section, we give the procedure for performing permutation tests based on MWMC, and analyze its validity (mainly the type II error). As a start, we introduce some notations and definitions.

Notations and Definitions

Let $\mathcal{S}_n$ be the group of permutations on n elements, we define the permutation operator $\sigma \in \mathcal{S}_n$ such that $\sigma(x_1, x_2, ..., x_n) \mapsto (x_{\sigma(1)}, x_{\sigma(2)}, ..., x_{\sigma(n)})$, where $\{x_1, x_2, ..., x_n\}$ is n

i.i.d. samples of random variable x. For example, when $n = 2, \sigma(1) = 2, \sigma(2) = 1$, then $\sigma(x_1, x_2) = (x_2, x_1)$. In this case σ represents the swap operation of two samples. Now, we define the statistic after permutation σ, which is easy to do by simply replacing the original sample (x, y) with the permuted sample $(\sigma(x), y)$. Along with the procedure in the last section, we have the following definitions.

Definition 4. (Permutation Empirical WMCov). *Let* $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ *be the set of n i.i.d. pairs drawn from the joint probability density function* $P_{x,y}$ *and* σ *be the permutation performed on the samples of x. Then, for levels i, j and shiftings m, k, the Permutation Empirical Wavelet Mapping Covariance is given by*

$$\widehat{Cov}^{\psi}_{i,j,mk}(\sigma x, y) = \frac{1}{n}\sum_{p}[\psi_i(x_{\sigma(p)} - m)\psi_j(y_p - k)] - \frac{1}{n(n-1)}\sum_{p\neq q}[\psi_i(x_{\sigma(p)} - m)\psi_j(y_q - k)]. \tag{7}$$

Definition 5. (Permutation Empirical MWMC). *Let* $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ *be the set of n i.i.d. pairs drawn from the joint probability density function* $P_{x,y}$ *and* σ *be the permutation performed on the samples of x. Then, for finite positive integer l, the Permutation Empirical MWMC is defined as*

$$\widehat{\rho}^{\psi}_{l}(\sigma x, y) := \sup_{0\le i,j\le l}\sup_{m,k\in\mathbb{Z}}\left|\frac{\widehat{Cov}^{\psi}_{i,j,mk}(\sigma x, y)}{\sqrt{\widehat{Var}^{\psi}_{i,m}(\sigma x) + \kappa}\sqrt{\widehat{Var}^{\psi}_{j,k}(y) + \kappa}}\right|, \tag{8}$$

where $\widehat{Var}^{\psi}_{i,m}(\sigma x) := \frac{1}{2n(n-1)}\sum_{p\neq q}[\psi_i(x_{\sigma(p)} - m) - \psi_i(x_{\sigma(q)} - m)]^2$.

Procedure: Now we describe the procedure of permutation tests using $\widehat{\rho}^{\psi}_{l}(\sigma x, y)$. First, we perform B permutation operations to obtain the sequence $\Theta := [\widehat{\rho}^{\psi}_{l}(x, y), \widehat{\rho}^{\psi}_{l}(\sigma_1 x, y), \cdots, \widehat{\rho}^{\psi}_{l}(\sigma_B x, y)]$, which contains $B + 1$ elements (including the statistic of observed data $\widehat{\rho}^{\psi}_{l}(x, y)$). Then, we calculate the one-side p-value $= \sum_{i=0}^{B}\mathbf{1}\{\widehat{\rho}^{\psi}_{l}(\sigma_i x, y) \ge \widehat{\rho}^{\psi}_{l}(x, y)\}/(B + 1)$, where we agree $\widehat{\rho}^{\psi}_{l}(\sigma_0 x, y) = \widehat{\rho}^{\psi}_{l}(x, y)$. If p-value $\le \alpha$, we reject the null hypothesis $\mathcal{H}_0 : x \perp\!\!\!\perp y$. More details are given in Algorithm 1.

Type I error: Under the null hypothesis $\mathcal{H}_0$, the sequence Θ is exchangeable, i.e., the joint probability distribution does not change when the positions of the elements in the sequence are altered. Intuitively, the observed data are equivalently obtained from the samples in the sequence. The exchangeability of sequence under $\mathcal{H}_0$ guarantees that the probability of type I error occurrence is no greater than the given significance level α. In the next, we discuss the bound on type II error.

Bound on Type II Error

In order to derive the bound on type II error, we need to obtain the distribution of the statistic $\widehat{\rho}^{\psi}_{l}(\sigma x, y)$ under the alternate hypothesis $\mathcal{H}_1$. We first estimate the range of Permutation Empirical MWMC by the following Lemma 2 and Lemma 3, whose proofs are given in Appendix.

Lemma 2. (Expectation of Permutation Empirical WMCov) *Assume each permutation of n elements has the same probability, then for fixed constants i, j, m, k,*

$$\mathbf{E}[\widehat{Cov}^{\psi}_{i,j,mk}(\sigma x, y)] = 0. \tag{9}$$

Lemma 3. (Variance of Permutation Empirical WMCov) (Variance of Permutation Empirical WMCov) *Assume each permutation has the same probability, then for fixed constants i, j, m, k, when* $n \ge 4$*,*

$$\mathbf{Var}[\widehat{Cov}^{\psi}_{i,j,mk}(\sigma x, y)] \le \frac{3}{n}. \tag{10}$$

Roughly speaking, Permutation Empirical WMCov is concentrated around 0 when the sample size is large enough. $\widehat{\rho}^{\psi}_{l}(\sigma x, y)$ inherits similar properties. We have the following Theorem 4:

Theorem 4. (Bound on Permutation Empirical MWMC). *Let* κ *be a positive constant and n be the sample size, then for* $n \ge 4$*, with probability at least* $1 - \delta$*,*

$$\widehat{\rho}^{\psi}_{l}(\sigma x, y) \le \sqrt{\frac{128/\delta}{\kappa^2}}\frac{2^l}{\sqrt{n}}. \tag{11}$$

Proof. The proof is given in Appendix.

Put all above together, we obtain the main theoretical result presented in Theorem 5:

Theorem 5. (Bound on Type II Error). *Let* (x, y) *be a L-lipschitz RV pair and each variable takes values in the interval* $[-1, 1]$*. Under the alternate hypothesis* $\mathcal{H}_1 : x \not\perp\!\!\!\perp y$*, i.e.,* $\rho^{\psi} > 0$*, let* $\rho^{\psi}_{l} = \rho_l$*,* $\alpha \in (0, 1)$ *be the significant level and finite positive constant B be the number of permutations, if* $\kappa < 0.1$*,* $n \ge 4$*,* $B \ge \frac{1}{\alpha} - 1$ *and* $l \ge \left\lceil \log_2(\frac{2(1+12L)}{\rho^{\psi}\kappa}) \right\rceil + 1$*, then*

$$P(\textit{Type II error}) \le \frac{2^{2l+9}B}{n\kappa^2\rho_l^2} + 4e^{-n\kappa^4\rho_l^2/24^2+2l+3}. \tag{12}$$

Proof. The proof is given in Appendix.

Therefore, when the sample size n is large enough, the type II error rate is asymptotically $O(n^{-1})$.

Algorithm

Here, we give the algorithmic details of permutation test using MWMC. For simplicity, our algorithm is called **W**avelet **I**ndependence **T**est (WIT), the process of WIT is outlined in Alg. 1. We first obtain the sequence $\Theta := [\widehat{\rho}^{\psi}_{l}(x, y), \widehat{\rho}^{\psi}_{l}(\sigma_1 x, y), \cdots, \widehat{\rho}^{\psi}_{l}(\sigma_B x, y)]$ by B permutations (Lines 1-18). The process of calculating $\widehat{\rho}^{\psi}_{l}(\sigma_t x, y)$ in Θ is divided into two main steps: first we generate the wavelet mappings of the data (Lines 5-9), and then calculate $\widehat{\rho}^{\psi}_{l}(\sigma_t x, y)$ (Lines 10-23) according to Eq. (8). After that, we calculate the p-value (Lines 26-30) to determine the independence.

Computational Complexity: In the typical setup (with a very large n, a large $B \ll n$ and a small $l \ll B$), the computational complexity of WIT is dominated by the process of calculating MWMC coefficient. Hence, the total time complexity is $O(B \cdot 4^l n) \approx O(n)$, and the space cost is $O(2^l n) \approx O(n)$.

Parallelizable implementation: In practice, we can parallelize the permutation step (Line 3-4). For the key step in the algorithm (Line 17-23), we can perform matrix operation optimization using GPU. Some results of empirical running time are given in the performance evaluation section.

Choice of Wavelet: Though our theoretical derivation applies to all compactly supported mother wavelets ϕ that can make Theorem 1 hold, the choice of ϕ is unavoidable. As we do not impose any assumption on the distribution, we cannot provide a concrete selection strategy. In our experiments, we choose linear B-spline as the mother wavelet due to its simple form and easy implementation.

Parameter Selection: In practice, we avoid the choice of κ by judging whether the variance exceeds a threshold to avoid a denominator of 0 (which is consistent with the purpose of adding κ). Since we observe that larger levels do not lead to significant improvement, we set level $l = 2$.

Algorithm 1: Wavelet Independence Test (WIT)

Input: observed data $x_{n\times 1}, y_{n\times 1}$, significance level α, permutation number B, wavelet level l.

Output: $X \perp\!\!\!\perp Y$ or $X \not\perp\!\!\!\perp Y$.

Initial permutation $\sigma_{n\times(B+1)} = [\sigma_0, \sigma_1, \sigma_2, ..., \sigma_B]$,
  $\widehat{\rho}_l^{\psi} = O_{1\times(B+1)}$
◁ **Parallel permutation step.**
for $\forall \sigma_t \in \sigma_{n\times(B+1)}$ **do**
  ◁ **Generate multi-level wavelet mapping of data.**
  $\Psi(\sigma_t x) \leftarrow [\psi_0(\sigma_t x - 1), \psi_0(\sigma_t x), \psi_0(\sigma_t x + 1), ...$
  $, \psi_l(\sigma_t x - 2^l), ..., \psi_l(\sigma_t x + 2^l)]_{n\times(2^{l+2}-1)}$
  $\Psi(y) \leftarrow [\psi_0(y - 1), \psi_0(y), \psi_0(y + 1), ..., \psi_l(y - 2^l),$
  $..., \psi_l(y + 2^l)]_{n\times(2^{l+2}-1)}$
  ◁ **Calculate MWMC coefficient.**
  Center every column of matrix to mean zero.
  $\Psi(\sigma_t x) \leftarrow \Psi(\sigma_t x) - \text{mean}(\Psi(\sigma_t x))$,
  $\Psi(y) \leftarrow \Psi(y) - \text{mean}(\Psi(y))$
  Calculate the sample variance.
  $\widehat{\text{Var}}^{\psi}(\sigma_t x) \leftarrow \text{sum}(\Psi(\sigma_t x) \odot \Psi(\sigma_t x))/(n-1)$,
  $\widehat{\text{Var}}^{\psi}(y) \leftarrow \text{sum}(\Psi(y) \odot \Psi(y))/(n-1)$
  for $i = 1 : 2^{l+2} - 1$ **do**
    $\psi_i \leftarrow$ the i^{th} column of $\Psi(\sigma_t x)$,
    $\Psi_{repeat} \leftarrow [\psi_i, \psi_i, ..., \psi_i]_{n\times(2^{l+2}-1)}$
    $\widehat{\text{Var}}^{\psi}[:, i] \leftarrow$ the i^{th} column of $\widehat{\text{Var}}^{\psi}(\sigma_t x)$
    $\widehat{\text{Cov}}^{\psi}(\sigma_t) \leftarrow \text{abs}(\text{sum}(\Psi_{repeat} \odot \Psi(y))/n(n-1))$
    $\widehat{\rho}_l^{\psi}(\sigma_t x, y) \leftarrow \max\Big\{\widehat{\text{Cov}}^{\psi}(\sigma_t)/(\widehat{\text{Var}}^{\psi}[:, i] + \kappa)^{\frac{1}{2}}$
    $/(\widehat{\text{Var}}^{\psi}(y) + \kappa)^{\frac{1}{2}}, \widehat{\rho}_l^{\psi}(\sigma_t x, y)\Big\}$
  end for
end for
Calculate p-value
  $p\text{-value} = \sum_{i=0}^{B} \mathbf{1}\{\widehat{\rho}_l^{\psi}(\sigma_i x, y) \geq \widehat{\rho}_l^{\psi}(\sigma_0 x, y)\}/(B+1)$.
if p-value $\leq \alpha$ **then**
  return $X \not\perp\!\!\!\perp Y$.
else
  return $X \perp\!\!\!\perp Y$.
end if

Performance Evaluation

In this section, we demonstrate the effectiveness of our method WIT through extensive experiments [3] and comparison with existing methods, including DARING (He et al. 2021), KCIT (Zhang et al. 2011), HSIC (Gretton et al. 2005a), FRCIT (Zhang et al. 2021), SCIT (Zhang et al. 2022) and RDC (Lopez-Paz, Hennig, and Schölkopf 2013). For comparing with RDC, we follow (Bellot and van der Schaar 2019) to implement RDC with permutation test, and we call it RDCPT hereafter. We first evaluate these methods on the independence test and conditional independence test by diverse synthetic data. We then conduct the experiment on a real dataset to infer causality. Finally, we analyze the consistency, efficiency, and sensitivity of the proposed method. Details of the implementation of our method and the existing methods are described in Appendix.

Independence

Setup: Here, we generate synthetic data according to the additive noise model as in (Ramsey 2014; Zhang et al. 2011). Define (x, y) under null hypothesis $\mathcal{H}_0$ and alternate hypothesis $\mathcal{H}_1$ as follows:

$$\begin{aligned} \mathcal{H}_0: &\ \begin{matrix} x = f(\epsilon_1 + \epsilon_2) + \epsilon_3 + \epsilon_4 \\ y = g(\epsilon_5 - \epsilon_6) + \epsilon_7 \end{matrix} \\ \mathcal{H}_1: &\ \begin{matrix} x = f(\epsilon_1 + \epsilon_2) + \epsilon_3 + \epsilon_4 \\ y = g(\epsilon_1 - \epsilon_2) + \epsilon_7 \end{matrix} \end{aligned}, \tag{13}$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_7$ have jointly independent distributions and f, g denote smooth functions. We choose ϵ_i with uniform distribution $U(-1/2, 1/2)$, chi-square distribution $\chi^2(1)$, beta distribution $\text{Be}(0.2, 0.4)$, Laplace distribution $\text{Lap}(0, 1)$, exponential distribution $\text{Exp}(1/2)$, and choose f, g from $\{(\cdot), \sin(\cdot), \exp(-\|\cdot\|), \log(\cdot)^2, (\cdot)^2\}$. In order to reflect the performance of nonlinear dependence of the methods, we design the synthetic data to make their correlation as close to zero as possible. At the same time, we want the signal-to-noise ratio to be close to 1 to reflect the noise resistance of the methods. Furthermore, we use the heteroskedasticity setting to further increase the diversity. More details of the experimental setup are described in Appendix.

Results: We investigate the performance of the independence test for sample sizes of $\{1000, 2000\}$, and the experimental results are obtained by averaging 100 replicate experiments. Due to space limit, more experimental results are presented in Appendix. The experimental results with 1000 samples are presented in Table 1, where the first five tables (a-e) show the type II error rates of each method under different distribution settings, and the last table (f) presents the corresponding type I error rates (averaging the experimental results over the five distribution settings). Note that we do not show the standard deviations because they are very small. It can be seen that the type I error rates of all methods are controlled around 0.05, and our method outperforms the other methods in terms of average type II error rate. An interesting finding is that not any method can beat all the

[3] The experimental source code and datasets are given in the repository https://github.com/renyixin666/WIT-Code.

Algorithm	linear	sin	exp	log	square	Average
DARING	1.00	1.00	0.96	**0.00**	**0.14**	0.62
KCIT	0.53	0.70	0.75	**0.00**	0.73	0.54
HSIC	0.43	0.66	0.68	**0.00**	0.53	0.46
FRCIT	**0.04**	**0.28**	0.90	0.57	0.92	0.66
RDCPT	0.30	0.52	0.51	**0.00**	0.26	0.32
SCIT	0.74	0.72	0.62	**0.00**	0.44	0.50
WIT	0.22	0.40	**0.46**	**0.00**	0.19	**0.25**

(a) Type II error rate on uniform distribution

Algorithm	linear	sin	exp	log	square	Average
DARING	0.41	0.46	1.00	0.99	0.03	0.58
KCIT	**0.00**	**0.00**	**0.00**	**0.00**	0.35	0.07
HSIC	**0.00**	**0.00**	**0.00**	**0.00**	0.38	0.08
FRCIT	0.01	0.04	0.97	0.95	0.50	0.49
RDCPT	**0.00**	**0.00**	**0.00**	**0.00**	0.21	0.04
SCIT	0.29	0.32	0.02	0.97	0.05	0.33
WIT	**0.00**	**0.00**	**0.00**	**0.01**	**0.02**	**0.01**

(b) Type II error rate on Chi-square distribution

Algorithm	linear	sin	exp	log	square	Average
DARING	0.99	0.99	0.63	**0.00**	0.04	0.53
KCIT	**0.00**	**0.00**	0.16	**0.00**	0.16	0.06
HSIC	**0.00**	**0.00**	0.27	**0.00**	0.10	0.07
FRCIT	0.09	0.34	0.89	0.64	0.75	0.54
RDCPT	**0.00**	**0.00**	**0.05**	**0.00**	**0.01**	**0.01**
SCIT	0.53	0.73	0.70	**0.00**	0.28	0.45
WIT	**0.00**	**0.00**	0.25	**0.00**	0.08	0.07

(c) Type II error rate on Beta distribution

Algorithm	linear	sin	exp	log	square	Average
DARING	0.57	0.60	1.00	**0.00**	0.86	0.61
KCIT	0.25	0.27	0.20	**0.00**	0.94	0.33
HSIC	0.18	0.21	0.09	**0.00**	0.93	0.28
FRCIT	0.20	0.24	0.91	0.50	0.93	0.55
RDCPT	0.11	0.10	0.09	**0.00**	0.80	0.22
SCIT	0.11	0.10	0.09	**0.00**	0.87	0.23
WIT	**0.06**	**0.04**	**0.07**	**0.00**	**0.78**	**0.19**

(d) Type II error rate on Laplace distribution

Algorithm	linear	sin	exp	log	square	Average
DARING	0.51	0.46	1.00	1.00	0.23	0.64
KCIT	**0.00**	**0.00**	0.22	0.13	**0.13**	0.10
HSIC	**0.00**	**0.00**	0.14	**0.00**	0.71	0.17
FRCIT	0.04	0.12	0.94	0.99	0.67	0.55
RDCPT	**0.00**	**0.00**	**0.01**	**0.00**	0.43	0.09
SCIT	0.46	0.44	0.35	0.94	0.21	0.48
WIT	**0.00**	**0.00**	**0.01**	0.13	**0.13**	**0.06**

(e) Type II error rate on exponential distribution

Algorithm	linear	sin	exp	log	square	Average
DARING	0.04	0.03	0.02	0.03	0.02	0.03
KCIT	0.06	0.06	0.05	0.06	0.03	0.05
HSIC	0.06	0.06	0.07	0.08	0.03	0.06
FRCIT	0.05	0.06	0.04	0.03	0.06	0.05
RDCPT	0.04	0.05	0.05	0.06	0.04	0.05
SCIT	0.04	0.05	0.03	0.04	0.03	0.04
WIT	0.04	0.03	0.04	0.04	0.04	0.04

(f) Average type I error rate on five distributions

Table 1: Experimental results of independence tests with 1000 samples

others in all cases, indicating that each method has its performance bias when the sample size is limited. In the case of Laplace-square, all methods perform poorly due to the fact that the square function $(\cdot)^2$ deteriorates the smoothness of the distribution and increases more noise than the Laplace-linear case. As for the log case, most of the methods perform better because the $\log(\cdot)^2$ function increases the proportion of the signal (e.g. function $\log(\cdot)^2$ map $[0.01, 1]$ to $[-9.2, 0]$).

Conditional Independence

Setup: Here we compare the performance of the methods on residual-based conditional independence test (Ramsey 2014). In the case of additive noise model ($x = f(z) + \epsilon_1, y = g(z) + \epsilon_2$), conditional independence is tested by converting $x \perp\!\!\!\perp y|z$ equivalently to $x - \mathbf{E}[x|z] \perp\!\!\!\perp y - \mathbf{E}[y|z]$. We define (x, y) as $x = c(\sum_{i=1}^{r} z_i/5) + \epsilon_1, y = c(\sum_{i=1}^{r} z_i/5) + \epsilon_2$, where $z_1, z_2, ..., z_5 \sim U(-1/2, 1/2)$ have joint independent distributions, ϵ_1, ϵ_2 are mutually independent standard Gaussian random variables and the constant $c = \{7.0, 6.0, 5.0, 4.0, 3.0\}$ is used to adjust the signal-to-noise ratio. We randomly choose a certain number (from 0 to 5) of z_i as the multiple regression variable R and then test whether $x \perp\!\!\!\perp y|R$. Here we use least squares regression for the calculation of the residual term. Details of the residual term calculation and experimental setup are described in Appendix.

Results: For visualization, we take the top-5 methods from the last experiment for comparison, and each data point is the average result of 500 experiments. The results of type II error rate with sample size $n = 250$ are shown in Fig. 1, while complete results are presented in Appendix. The type I error is controlled around 0.05 for all methods. It can be seen that WIT achieves the lowest type II error rate in the case of regression variables 0-3 with $c = \{7.0, 6.0, 5.0\}$. These results indicate that WIT is generally more robust to noise. As the signal-to-noise ratio continues to decrease, all methods become unreliable given 4 regression variables, at this time their type II error rates are all close to 1.

Causal Discovery from Real Data

Here we evaluate all methods on causal skeleton learning, as the resulting skeletons are completely determined by independence test. We use the well-known Saches (Sachs et al. 2005) dataset for experiment, which is a real causal protein signaling network that measures protein expression levels and is widely used in causal discovery tasks (Ng, Ghassami, and Zhang 2020; He et al. 2021). The dataset contains a total of 853 samples and a corresponding causal graph [4] (11 nodes and 18 arcs) that is usually considered as the ground truth. The results in Table 2, from which we can see that our method outperforms the other methods as a whole in terms of average performance F1-score.

[4]The causal graph of Saches is given in https://www.bnlearn.com/bnrepository/.

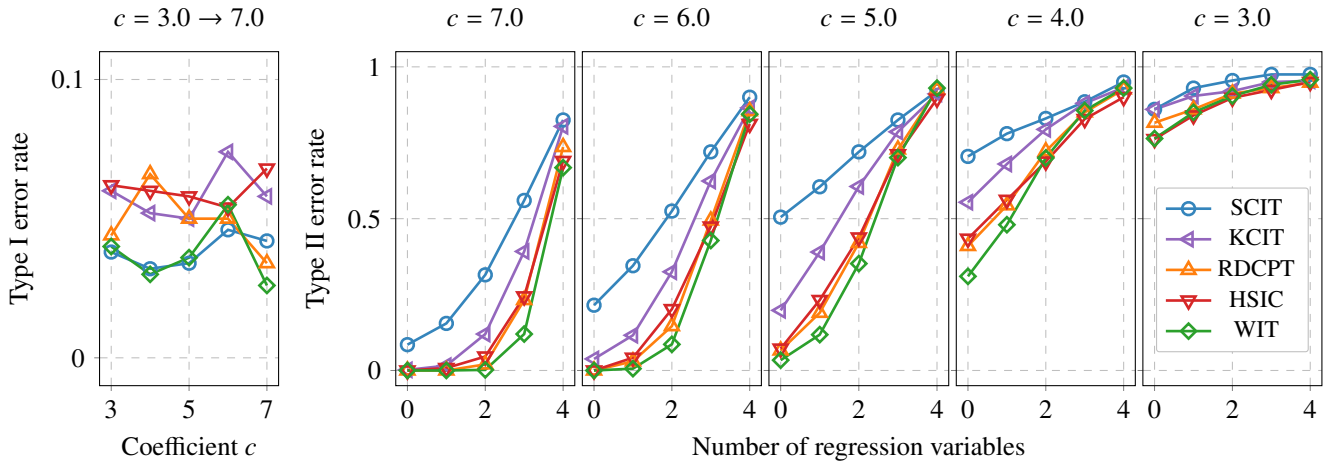

Figure 1: Experimental results of two types of errors in conditional independence tests, with sample size $n = 250$. Lower is better. The left column shows the relationship between the type I error and the coefficient c for each method. Each column on the right shows the relationship between the type II error rate and the number of regression variables for different values of coefficient c. The error bars are not shown in the figures to make all the curves discernable.

Algorithm	Recall	Precision	F1-score
DARING	0.4706	**1.0000**	0.6400
KCIT	**0.5294**	0.8182	0.6429
HSIC	**0.5294**	0.7193	0.6096
FRCIT	0.1824	0.7757	0.2927
RDCPT	**0.5294**	0.7500	0.6207
SCIT	0.2006	0.5890	0.2960
WIT	**0.5294**	0.9228	**0.6719**

Table 2: Results on the Sachs dataset.

Efficiency, Consistency and Sensitivity

Efficiency: We compare the efficiency of the methods. We repeat the experiment 50 times with sample size of 100, 500, 1000, 5000, 10,000, 100,000 and 1,000,000. The experimental results are averaged over 50 replicate experiments. All experiments are conducted on a PC with a NVIDIA RTX 2080Ti GPU, eight 3.00 GHz CPU cores and 32GB RAM.

Here, we focus on four methods KCIT, HSIC, RDCPT and WIT as they outperform the other methods in most cases. The results are presented in Appendix, from which we can see that WIT works much more efficiently than RDCPT, especially for the sample sizes from 100 to 1000. And compared to the kernel-based methods, WIT is more efficient when sample size is greater than 5,000. When sample size is larger than 1,00,000, the running time of WIT gradually shows a linear growth trend, which is consistent with the time complexity of the algorithm.

Consistency: We explore the consistency (the type II error rate tends to zero as the sample size increases) of our method. We keep the same experimental setup as in the independence test experiment and test the type II error rate for each case with the sample size $n = \{500, 1000, 1500, 2000, 2500\}$. For the Laplace-square case, the type II error rate remains high at the sample size of 2500, so we further stepwisely increase the sample size to $50,000$. The complete experimental results are presented in Appendix. We can see that as the sample size increases, the type II error rate gradually decreases and finally approaches 0, which validates the consistency of our method, as stated by Theorem 5.

Sensitivity: We check the sensitivity of our method to the number of permutations B. We take the uniform-linear case for experiment, and the setting remains the same as before. The experimental results for different permutations $B = \{100, 200, 300, 400, 500\}$ are shown in Appendix. It can be seen that the variation of B between 100 and 500 has no significant effect on the two types of error rates, indicating the robustness of our method with respect to B.

Conclusion

In this work, we present a novel criterion for measuring nonlinear dependence between two real variables by estimating their correlation under different levels of wavelet mappings. We also design a permutation test based on this criterion for independence test. We provide theoretical guarantees for this test, ensuring that the type II error rate decreases rapidly as the sample size increases, while keeping the type I error rate controllable. Experiments on various data show that compared to the existing methods, our method i) is more effective for different distributions of data, ii) is generally more robust to noise and thus can handle better the data of smaller signal-to-noise ratio, iii) can be applied to causal discovery and achieve better performance. Future works include constructing independence testing algorithms based on other mother wavelets according to our proposed procedure, optimizing the computation time of the criterion, and studying the sensitivity of the number of permutations to different distribution settings.

Acknowledgements

This work was supported by National Key Research and Development Program of China (grant No. 2021YFC3340302), and partially by National Natural Science Foundation (NSFC) (U1936205, 61972100 and 62006051). Hao Zhang was also supported by China Postdoctoral Science Foundation (2022M720033).

References

Bach, F. R.; and Jordan, M. I. 2002. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul): 1–48.

Bellot, A.; and van der Schaar, M. 2019. Conditional independence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32.

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, 1–4. Springer.

Berlinet, A.; and Thomas-Agnan, C. 2011. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Casella, G.; and Berger, R. L. 2021. *Statistical inference*. Cengage Learning.

Chatterjee, S. 2021. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536): 2009–2022.

Comon, P. 1994. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314.

Daubechies, I. 1992. *Ten lectures on wavelets*. SIAM.

Daudin, J. 1980. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3): 581–590.

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

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

Gretton, A.; Herbrich, R.; and Smola, A. J. 2003. The kernel mutual information. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 4, IV–880. IEEE.

Gretton, A.; Smola, A.; Bousquet, O.; Herbrich, R.; Belitski, A.; Augath, M.; Murayama, Y.; Pauls, J.; Schölkopf, B.; and Logothetis, N. 2005b. Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, 112–119. PMLR.

He, Y.; Cui, P.; Shen, Z.; Xu, R.; Liu, F.; and Jiang, Y. 2021. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 596–605.

Heller, R.; Heller, Y.; and Gorfine, M. 2013. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2): 503–510.

Hoeffding, W. 1948. A non-parametric test of independence. *The annals of mathematical statistics*, 546–557.

Lopez-Paz, D.; Hennig, P.; and Schölkopf, B. 2013. The randomized dependence coefficient. *Advances in neural information processing systems*, 26.

Lyons, R. 2013. Distance covariance in metric spaces. *The Annals of Probability*, 41(5): 3284–3305.

Muller, K.-R.; Mika, S.; Ratsch, G.; Tsuda, K.; and Scholkopf, B. 2001. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2): 181–201.

Ng, I.; Ghassami, A.; and Zhang, K. 2020. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33: 17943–17954.

Pavlov, A. N.; Hramov, A. E.; Koronovskii, A. A.; Sitnikova, E. Y.; Makarov, V. A.; and Ovchinnikov, A. A. 2012. Wavelet analysis in neurodynamics. *Physics-Uspekhi*, 55(9): 845.

Plackett, R. L. 1983. Karl Pearson and the chi-squared test. *International statistical review/revue internationale de statistique*, 59–72.

Qi, G.-J. 2020. Loss-sensitive generative adversarial networks on lipschitz densities. *International Journal of Computer Vision*, 128(5): 1118–1140.

Ramsey, J. D. 2014. A scalable conditional independence test for nonlinear, non-Gaussian data. *arXiv preprint arXiv:1401.5031*.

Rényi, A. 1959. On measures of dependence. *Acta mathematica hungarica*, 10(3-4): 441–451.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529.

Schweizer, B.; and Wolff, E. F. 1981. On nonparametric measures of dependence for random variables. *The annals of statistics*, 9(4): 879–885.

Székely, G. J.; and Rizzo, M. L. 2009. Brownian distance covariance. *The annals of applied statistics*, 3(4): 1236–1265.

Torrence, C.; and Compo, G. P. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1): 61–78.

Zhang, H.; Zhang, K.; Zhou, S.; Guan, J.; and Zhang, J. 2021. Testing Independence Between Linear Combinations for Causal Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6538–6546.

Zhang, H.; Zhou, S.; Zhang, K.; and Guan, J. 2022. Residual Similarity Based Conditional Independence Test and Its Application in Causal Discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5942–5949.

Zhang, K. 2019. BET on Independence. *Journal of the American Statistical Association*, 114(528): 1620–1637.

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Kernel-based conditional independence test and application in causal discovery. 804–813. Corvallis, OR.