On nonparametric conditional independence tests for continuous variables

Chun Li | Xiaodan Fan

Faculty of Science, Tianjin University of Technology and Education, Tianjin, China

Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong

Correspondence
Xiaodan Fan, Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong.
Email: xfan@cuhk.edu.hk

Funding information
NSF of China, Grant/Award Number: Tianyuan fund for Mathematics ZZG0011621; NSF of Tianjin, Grant/Award Number: Young Scientists Fund KYQD1672; Research Grants Council of the Hong Kong Special Administrative Region, China, Grant/Award Numbers: General Research Fund 14203915, 14173817; TUTE grant, Grant/Award Number: 18JCQNJC70100

Abstract
Testing conditional independence (CI) for continuous variables is a fundamental but challenging task in statistics. Many tests for this task are developed and used increasingly widely by data analysts. This article reviews the current status of the nonparametric part of these tests, which assumes no parametric form for the joint continuous density function. The different ways to approach the CI are summarized. Tests are also grouped according to their data assumptions and method types. A numerical comparison is also conducted for representative tests.

This article is categorized under:
Statistical and Graphical Methods of Data Analysis > Analysis of High Dimensional Data
Statistical and Graphical Methods of Data Analysis > Multivariate Analysis

KEYWORDS
conditional independence, hypothesis testing, literature review

1 | INTRODUCTION

Independence and conditional independence (CI) are fundamental concepts in statistics. Dawid (1979) and Dawid (1980) discuss important properties of CI for exploring queries in statistical inference, such as sufficiency, parameter identification, adequacy, and ancillarity. Blitzstein (2011) states that conditioning is the soul of statistics. With the rise of the related application fields including causal inference, graphical model learning and feature selection (Koller & Sahami, 1996; Xing, Jordan, & Karp, 2001), CI testing has attracted increasing attention from both theoretical side and application side.

The high demand of CI tests is related to the common desire to identify the relationships among different objects, including how they are connected, what is the mechanism they interact with each other, and how information flows among these objects (Gerstein et al., 2012). This desire entails the attraction of probabilistic graphical models, which express the relations among variables in terms of independence and CI. They are now extremely widely used in various disciplines, for example, information extraction, speech recognition, computer vision, modeling of gene regulatory networks, gene finding, and diagnosis of diseases. As the building blocks of graphical models, CI relations are the keys to answer scientific questions of these areas.

Formally, the problem of interest is testing the CI between two random variables $X$ and $Y$ given a third random vector $Z$ based on a finite sample. In statistical symbol, the null hypothesis is written as $H_0 : X \perp Y \mid Z$, where $\perp$ denotes “independent from.” The alternative hypothesis is written as $H_1 : X \nparallel Y \mid Z$, where $\nparallel$ denotes “dependent with.”
Conditional independence tests for the case in which the variables are categorical are abundant in literature. Most well-known tests include the Pearson's \( \chi^2 \) test and the likelihood ratio test (Edwards, 2000; Pearson, 1992; Tsamardinos & Boubadakis, 2010). If \( Z \) is discrete, CI test turns into testing for unconditional independence \( X \perp Y | Z = z \) for every possible value \( z \) that \( Z \) takes (Tsamardinos & Boubadakis, 2010).

Testing CI of continuous random variables is a much more difficult case (Bergsma, 2004). That prompts the use of simplifying assumptions. In the early days, one commonly used method is the test based on the partial correlation coefficient, which requires the assumption of Gaussian variables (Fisher, 1915, 1921). Milder assumptions such as linearity and additivity are well studied in the literature (Daudin, 1980; Spirtes, Glymour, & Scheines, 1993; Geiger & Heckerman, 1994; Peters, Mooij, Janzing, & Schölkopf, 2014). However, real-life situations where linear and additive conditions are not met are still plenty (e.g., physiological signals, stock market prices, weather status, etc.; Margaritis, 2005), thus recent research has focused on nonparametric CI tests, where no assumption on the functional form between the variables or the data distributions is imposed. However, nonparametric CI testing faces many challenges. In general, current nonparametric CI tests either suffer from the “curse of dimensionality” in terms of the dimensionality of the conditioning vector \( Z \) (Bergsma, 2004), or are limited by scalability, that is, costing too much time in the large sample size setting. Moreover, Shah and Peters (2019) prove that no CI test can control Type-I error for all CI cases. More specifically, if the joint distribution of \((X, Y, Z)\) is continuous, a CI test that both (a) controls Type-I error under the significance level \( \alpha \) over the entire class of distributions satisfying CI, and (b) has power better than the significance level \( \alpha \) against any alternative hypothesis, cannot exist.

Although the dream of nonparametric CI tests to be both uniformly valid and powerful may not be achievable (Shah & Peters, 2019), their validity for a wide range of CI cases and their robust high power against a wide range of alternative hypotheses still make them the popular choice for many CI test problems. But since there is no non-trivial CI test that is valid for all CI distributions, it is crucial to choose the right test. Thus it is extremely important to understand these nonparametric CI tests, as well as understand the specific problem in hand. For the latter task, domain knowledge for the specific problem is needed. For the former task, we write this article to provide a brief review of current nonparametric CI tests for continuous variables, with emphasis on the CI characterization they are based on and the methodology line they used. We also provide an empirical comparison to show their performances on a range of popular distributions.

### 2 | CI CHARACTERIZATION APPROACHES

Let \( X, Y, \) and \( Z \) be three random vectors with domains \( X, Y, \) and \( Z, \) respectively, and assume that their joint distribution is continuous with density \( p_{X,Y,Z}(x,y,z) \). By definition, the CI between \( X \) and \( Y \) given \( Z \) is defined as \( p_{X,Y|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z) \). Obviously, it is impossible to check this equality for every value of \((x, y, z)\) based on a finite sample. Thus, researchers have come up with different approaches to characterize the CI. Some CI characterizations are of weaker sense (Daudin, 1980) as compared to the above CI definition, thus we call the characterizations equivalent to the above CI definition as “strong” CI. The word “strong” is only added to distinguish from the definition of weak CI (Daudin, 1980). CI or “strong” CI implies weak CI, but the converse is not true.

#### 2.1 | “Strong” CI

There are at least the following five approaches to characterize CI.

(A) We say that \( X \) is conditionally independent of \( Y \) given \( Z \), that is, \( X \perp Y | Z \), if the joint distribution \( p_{X,Y,Z}(x,y,z) \) factorizes as \( p_{X,Y,Z}(x,y,z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z) \) (or equivalently, \( p_{X,Y|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z) \)).

(B) Daudin (1980) provides a characterization of CI by enforcing the uncorrelatedness of functions in certain \( L^2 \) spaces. Let \( L^2_{X,Z} \) denote the space of all functions \( f \) such that \( \mathbb{E} f(X,Z)^2 < \infty \), that is, \( L^2_{X,Z} = \{ f | \mathbb{E} f(X,Z)^2 < \infty \} \), and define \( L^2_{XZ}, L^2_{X}, L^2_{Z} \) analogously. Consider the constrained \( L^2 \) spaces

\[
\mathcal{F}_{XZ} \triangleq \{ f \in L^2_{X,Z} | \mathbb{E}[ f(X,Z)|Z] = 0 \},
\]
\[ \mathcal{F}_{YZ} \triangleq \{ \tilde{g} \in L^2_{YZ} | \mathbb{E}[\tilde{g}(Y,Z) | Z] = 0 \} . \]

\[ \mathcal{F}_{Y|Z} \triangleq \{ \tilde{h} | \tilde{h} = h(Y) - \mathbb{E}[h'(Y) | Z] = 0, \tilde{h} \in L^2_Y \} . \]

The following conditions are equivalent:

(i) \( X \perp Y | Z \),

(ii) \( \mathbb{E} f(X,Z) \tilde{g}(Y,Z) = 0, \forall \tilde{f} \in \mathcal{F}_{XZ} \) and \( \forall \tilde{g} \in \mathcal{F}_{YZ} \),

(iii) \( \mathbb{E} f(X,Z) \tilde{h} (Y,Z) = 0, \forall \tilde{f} \in \mathcal{F}_{XZ} \) and \( \forall \tilde{h} \in \mathcal{F}_{Y|Z} \).

Note: The above result can be viewed as an extension of the partial correlation-based characterization of CI for Gaussian variables. For one-dimensional variables \( X \) and \( Y \), suppose that \( (X, Y, Z) \) is jointly Gaussian, then the partial correlation coefficient \( \rho_{X, YZ} \) (the correlation between the residuals \( e_X \) and \( e_Y \) resulting from the linear regression of \( X \) with \( Z \) and of \( Y \) with \( Z \), respectively) equals 0 if and only if \( X \perp Y | Z \). As in the above characterization (B), if we construct functions \( \tilde{f} \) and \( \tilde{g} \) from the corresponding \( L^2 \) spaces through nonlinear regression, for example, for any function \( f \in L^2_{XZ} \), define \( \tilde{f} \) by \( \tilde{f}(X,Z) = f(X,Z) - \mathbb{E}[f(X,Z) | Z] = f(X,Z) - h_j'(Z) \), where \( h_j'(Z) \in L^2_Z \) is the regression function of \( f(X,Z) \) on \( Z \), then the claim (ii) implies that \( X \perp Y | Z \) is equivalent to the vanishing of the correlation between any “residual” function of \( (X, Z) \) given \( Z \) and that of \( (Y, Z) \) given \( Z \).

(C) Characterization of CI can also be given in terms of the cross-covariance operator on reproducing kernel Hilbert spaces (RKHSs; Fukumizu, Gretton, Sun, & Schölkopf, 2008). Let \( k_X \) be a positive definite kernel on \( \mathcal{X} \), and \( \mathcal{H}_X \) be its corresponding RKHS. Define \( k_Y, \mathcal{H}_Y, k_Z, \) and \( \mathcal{H}_Z \) similarly. The cross-covariance operator \( \Sigma_{XY} : \mathcal{H}_Y \rightarrow \mathcal{H}_X \) is defined as follows:

\[ \langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_X} = \mathbb{E}_{XY} [f(X)g(Y)] - \mathbb{E}_X [f(X)] \mathbb{E}_Y [g(Y)] , \]

for all \( f \in \mathcal{H}_X \) and \( g \in \mathcal{H}_Y \). Thereby the definition of the conditional cross-covariance operator of \( (X, Y) \) given \( Z \) can be presented by

\[ \Sigma_{XY|Z} = \Sigma_{XY} - \Sigma_{XZ} \Sigma_{ZZ}^{-1} \Sigma_{ZY} . \]

Now let \( \tilde{X} = (X,Z) \), \( k_{\tilde{X}} = k_X k_Z \), and \( \mathcal{H}_{\tilde{X}} \) be the RKHS corresponding to \( k_{\tilde{X}} \). Assume \( \mathcal{H}_X \subseteq L^2_{XZ} \), \( \mathcal{H}_Y \subseteq L^2_{YZ} \), and \( \mathcal{H}_Z \subseteq L^2_{Z} \). Further assume that \( k_X k_Y \) is a characteristic kernel on \((\mathcal{X} \times \mathcal{Y}) \times Z\), and that the direct sum of \( \mathcal{H}_Z \) and \( \mathbb{R} \) is dense in \( L^2_Z \). Then \( X \perp Y | Z \) if and only if

\[ \Sigma_{XY|Z} = 0 . \]

Here a kernel \( k_X \) is considered as characteristic if \( \mathbb{E}_{X|Y} [f(X)] = \mathbb{E}_{X|Z} [f(X)] \), \( \forall f \in \mathcal{H}_X \) implies \( \mathbb{P}_X = \mathbb{Q}_X \), where \( \mathbb{P}_X \) and \( \mathbb{Q}_X \) are two probability distributions of \( X \).

Note: In K. Zhang, Peters, Janzing, and Schölkopf (2012), the authors point out that though the characterization (B) is intuitively more appealing, the characterization (C) is more useful, since in (B) one has to consider all functions in \( L^2 \), whereas (C) exploits the RKHSs corresponding to some characteristic kernels, which usually are much smaller. In fact, if we restrict the functions \( f \) and \( h' \) in claim (ii) of (B) to the space \( \mathcal{H}_{\tilde{X}} \) and \( \mathcal{H}_Z \), respectively, then (B) is reduced to (C).

(D) Conditional mutual information (CMI) can be applied to characterize CI. CMI for random variables \( X, Y, \) and \( Z \) is defined as
From the definition of CMI it is immediately clear that $X \perp Y \mid Z$ if and only if $I(X; Y \mid Z) = 0$.

2.2 | Weak CI

Daudin (1980) defined the weak CI, which can be considered as a mean property of CI.

(E) Following the expression used in Q. Zhang et al. (2017), the weak CI between $X$ and $Y$ given $Z$ is characterized by zero expected conditional covariances of square integrable functions, that is,

$$0 = \mathbb{E}_Z [\text{Cov}[f(X), g(Y) | Z]]$$

$$= \mathbb{E}_{XYZ} [(f(X) - \mathbb{E} f(X | Z))(g(Y) - \mathbb{E} g(Y | Z))]$$

$$= \mathbb{E}_{XYZ} [n_f(X,Z)n_g(Y,Z)],$$

for all $f \in L^2_X$ and $g \in L^2_Y$, where $n_f(X,Z) = f(X) - \mathbb{E} f(X | Z)$ and $n_g(Y,Z) = g(Y) - \mathbb{E} g(Y | Z)$ are the regression residuals for $f(X)$ and $g(Y)$, respectively.

We can see from above that weak CI is equivalent to the uncorrelatedness of residuals $n_f(X,Z)$ and $n_g(Y,Z)$ for all $f \in L^2_X$ and $g \in L^2_Y$. Obviously, CI implies weak CI but the converse is not true. Though tests based on weak CI cannot fully characterize CI, they potentially benefit from a simpler procedure and an improved power.

3 | SPECIFIC CI TEST GROUPS

In this section, we give a brief overview of some state-of-the-art nonparametric CI testing methods proposed in the literature according to their different data assumption and methodologies.

3.1 | Tests with IID data

3.1.1 | Discretization-based tests

This type of tests discretizes the domain of variable $Z$ as $\{z_1, z_2, \ldots, z_k\}$ in some optimal fashion, resulting in a partition of the data set into $k$ disjoint data sets, and hereby transform CI to the unconditional ones, $X \perp Y | Z = z_i, \forall i$. This idea resembles the characterization of CI when $Z$ is discrete. For instance, Margaritis (2005) obtains a posterior probability of independence from each subset; the test in T.M. Huang (2010) is constructed using the estimator of a weighted average of the form

$$\sum_{i=1}^{k} p_{Z}(z_i)p_{ii}(X,Y|Z = z_i),$$

where $p_{ii}(X,Y|Z)$ is the maximal nonlinear conditional correlation of $X$ and $Y$ given $Z$. Inevitably, discretization suffers from the curse of dimensionality, as high-dimensional conditioning set $Z$ requires a dramatically large sample size (Muandet, Fukumizu, Sriperumbudur, & Schölkopf, 2017). On the other hand, this type of CI tests are expected to work well if the joint distribution of $(X, Y, Z)$ can be approximated well by a certain discretization of $Z$.

3.1.2 | Metric-based tests

The intuition behind this category is the observation in (A) and extending measures of marginal independence to the conditional setting. For example, Su and White (2008) and M. Huang, Sun, and White (2016) use the weighted Hellinger distance and a family of generically comprehensively revealing functions, respectively, to measure the distance between $p_{X,Y,Z}p_Z$ and $p_{X,Y,Z}$. 

$$I(X; Y | Z) = \int_x \int_y \int_z \log \left( \frac{p_{Z}(z)p_{X,Y,Z}(x, y, z)}{p_{X,Z}(x, z)p_{Y,Z}(y, z)} \right) p_{X,Y,Z}(x, y, z) dx dy dz$$

$$= \int_x \int_y \int_z \log \left( \frac{p_{X,Y|Z}(x, y | z)}{p_{X|Z}(x | z)p_{Y|Z}(y | z)} \right) p_{X,Y,Z}(x, y, z) dx dy dz.$$
Su and White (2007) and Su and White (2014) measure the distance between $p_{Y|X,Z}$ and $p_{YZ}$ by applying their conditional characteristic functions and the smoothed empirical likelihood ratio. The conditional distance independence test (CDIT) proposed by Wang, Pan, Hu, Tian, and Zhang (2015) is also based on conditional characteristic functions, like the method in Su and White (2007), whereas CDIT measures the distance between $p_{X,YZ}$ and $p_{XY}p_{YZ}$. Comparing to Su and White (2007)'s, the definition of CI used by CDIT can be dealt with and computed more easily. Runge (2018) derives the CMI test based on the result in (D), which is in fact a property of the Kullback-Leibler divergence from $p_{X,YZ}$ to $p_{XY}p_{YZ}$. Through a nearest-neighbor approach, the test efficiently adapts also to non-smooth distributions due to strongly nonlinear dependencies.

However, this category involves the explicit estimation of the conditional densities or their associated variants, which is usually not an easy task, hence the power of these tests also deteriorates rapidly when encountering the curse of dimensionality (K. Zhang et al., 2012).

### 3.1.3 Permutation-based two-sample tests

Tests in this category also resort to the observation in (A), but instead of measuring the distance between $p_{X,YZ}$ and $p_{X,YZ}^\text{CI}$, they generate samples from $p_{X,YZ}^\text{CI}$ and compare with the original data set, which is sampled from the true joint distribution $p_{X,YZ}$. For instance, Doran, Muandet, Zhang, and Scholkopf (2014) propose a kernel CI permutation test (KCIPT). They use a specific permutation of the samples to simulate data from $p_{X,YZ}^\text{CI}$, which unfortunately requires solving a time-consuming linear program, then perform a kernel-based two-sample test (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2012). Lee and Honavar (2017) observe that by increasing the number of bootstraps, KCIPT suffers from a loss of calibratedness although obtaining a higher power. The authors then propose the self-discrepancy CI test, with a parameter-free (for a given choice of kernel and distortion measure) test statistic based on modified unbiased estimate of maximum mean discrepancy (Gretton et al., 2012), which is the largest difference in empirical expectations over features of the given sample and its permuted counterpart in the unit ball of a RKHS. Sen, Suresh, Shanmugam, Dimakis, and Shakkottai (2017) first use a nearest-neighbor bootstrap to generate data from $p_{X,YZ}^\text{CI}$, and then apply binary classifiers to achieve the CI testing goal. As these tests essentially have the flavor of nearest neighbor by using distance metrics which weight each variable equally, the required sample size increases dramatically when the conditioning set becomes larger (Aggarwal, Hinneburg, & Keim, 2001).

### 3.1.4 Kernel-based tests

A Hilbert space embedding of a distribution, that is, a kernel mean embedding, has recently emerged as a powerful tool for statistical inference. See Muandet et al. (2017) for a review. Kernel-based methods in general have strong empirical performance in the presence of curse of dimensionality and estimating the kernel mean embedding is easier than estimating the distribution itself (Muandet et al., 2017).

The Hilbert-Schmidt independence criterion (HSIC) is initially proposed for testing unconditional independence. It can be interpreted in two ways: (a) a measure of the distance in the RKHS between the embedding of the joint distribution of $X$ and $Y$, $p_{X,Y}$, and the embedding of the product of $X$ and $Y$'s respective marginal distribution, $p_{X}p_{Y}$; (b) the Hilbert-Schmidt norm of the cross-covariance operator of $(X,Y)$. Fukumizu et al. (2008) extends HSIC to the CI setting through utilizing the Hilbert-Schmidt norm of the conditional cross-covariance operator, and the corresponding CI characterization is concluded in (C). However, the null distribution of the test statistics in Fukumizu et al. (2008) is unknown, and its approximation relies on a computationally expensive resampling scheme.

An excellent representative of this test category is the kernel CI test (KCIT), which is proposed by K. Zhang et al. (2012). KCIT is motivated by the characterizations of CI in (B). As depicted in the previous section, (C) can be viewed as a relaxation of (B) by using a smaller yet sufficiently rich class of functions from certain RKHSs. Intuitively, KCIT works by testing for vanishing correlation between residual functions in RKHSs. A prominent advantage of KCIT is that the null distribution of the test statistic is derived and can be estimated efficiently. Nonetheless, as KCIT involves computing matrix inverses which scale strictly greater than $O(N^2)$, KCIT suffers from the drawback of lack in scalability, which places it out of rage of large sample size analysis, and therefore high-dimensional scenes.

To remedy the aforementioned issue, Strobl, Zhang, and Visweswaran (2019) utilize random Fourier features to approximate KCIT and thereby proposes two tests, namely the randomized conditional independence test (RCIT) and the randomized conditional correlation test (RCoT). RCIT explores the partial cross-covariance matrix of $(X,Y)$ and $Y$, and RCoT $X$ and $Y$ after subjecting the variable sets to the nonlinear transformations and then nonlinearly regressing out the effect of $Z$. Simulation
shows that both RCIT and RCoT give similar performance as KCIT but they are orders of magnitude faster in large-scale settings. RCoT is also related to the kernelized two-step CI test in H. Zhang, Zhou, et al. (2017), which we will discuss later.

### 3.1.5 Regression-based tests

In causal discovery, the observed data are often assumed to be generated under the additive noise model (ANM): \( Y = f(X) + \epsilon \), where \( \epsilon \) is a random noise term that is independent of \( X \) (Hoyer, Janzing, Mooij, Peters, & Schölkopf, 2009; J. Peters, Janzing, & Scholkopf, 2011; Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006). So, CI in ANM can be characterized by the following way (H. Zhang et al., 2017): assume \( X \) and \( Y \) are some deterministic functions of \( Z \) plus an additive noise term, that is,

\[
X = f(Z) + n_x,
Y = g(Z) + n_y,
\]

where \( n_x \) and \( n_y \) are zero-mean random variables independent of \( Z \), then we have

\[
X \perp Y | Z \Leftrightarrow n_x \perp n_y.
\]

Therefore, tests in this category (Hoyer et al., 2009; Jonas Peters et al., 2014; H. Zhang, Zhou, & Guan, 2018) proceed mainly in two steps: (a) regressing \( X \) on \( Z \) and \( Y \) on \( Z \), and then (b) testing for the unconditional independence between the corresponding fitted residuals \( \hat{\epsilon}_x = X - \hat{E}[X|Z] \) and \( \hat{\epsilon}_y = Y - \hat{E}[Y|Z] \). However, if the data departs from the ANM assumption, which is not straightforward to verify, the two-step procedures can lead to a substantial inflation of false discoveries (H. Zhang et al., 2017). To address this problem, H. Zhang et al. (2017) propose the KRESIT test, where feature representations of \( X \) and \( Y \) are constructed before doing regressions in the first step. They explained that the feature representation of the responses in regression is their key strategy, as it greatly relaxes the modeling assumptions. The authors also point out that KRESIT is closely related to RCoT (Strobl et al., 2019), as these two tests essentially compute the same test statistic, but uses different estimation procedures of the asymptotic null distribution. Note that the two-step tests are effectively testing for weak CI defined in the characterization (E), since they focus on individual effects of the conditioning variable \( Z \) on \( X \) and on \( Y \). With the ability to incorporate nonlinearities, this category can also be considered as extensions of partial correlation.

H. Zhang et al. (2017) proposes a CI test to relax the test of \( X \perp Y \mid Z \) to simpler unconditional independence tests of \( X - f(Z) \perp Y - g(Z) \) and \( X - f(Z) \perp Z \) or \( Y - g(Z) \perp Z \) under the ANM assumption. The residual prediction test in Heinze-Deml, Peters, and Meinshausen (2018) performs a regression from \( X \) on \( Z \), using an appropriate basis expansion (if not available, use random features as an approximation), and then tests for dependence between the scaled residuals and \( Y, Z \).

Recently, Shah and Peters (2019) propose the generalized covariance measure (GCM) test. For univariate \( X \) and \( Y \), instead of tests for (unconditional) independence between the residuals from regressing \( X \) and \( Y \) on \( Z \), the GCM tests for vanishing correlation, which is a normalized covariance between the residuals. The GCM test can also extend to handle settings where \( X \) and \( Y \) are potentially high dimensional. Empirical comparison shows that the GCM test performs well when ANM assumption is violated.

Recall the aforementioned argument in Shah and Peters (2019), there is no universally valid CI test, and domain knowledge should be considered for finding an appropriate test for certain data. However, obtaining knowledge on the entire joint distribution for seeking the suitable test is usually challenging. In this context, regression-based tests have an extra advantage, in that they convert the problem of guessing the whole joint distribution to selecting regression procedures, which is a more familiar task for both practitioners and researchers. Note that the validity of GCM is proven to be relied almost entirely on the weak requirement that the regression methods can estimate the conditional means \( X \) given \( Z \), and \( Y \) given \( Z \), at a slow rate (Shah & Peters, 2019).

### 3.1.6 Other tests

There exist surely many other CI tests based on IID data. For example, Song (2009) provides a slightly weaker test, which tests the hypothesis \( X \perp Y \mid \lambda_\theta(Z) \), where \( \lambda_\theta() \) is a scalar-valued function known up to an estimable parameter \( \theta \). Bergsma (2010) derives a CI test based on a new concept, the partial copula.
3.2 Tests with non-IID data

Granger causality, proposed in Granger (1969), has been widely used in economics since the 1960s and gains growing interest in neuroscience over recent years. The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another, which is a form of CI (Florens & Fougere, 1996). Most of the existing CI tests are derived under an IID assumption. When encountering non-IID data, these tests will all be biased. There are a few nonparametric CI tests designed for time series data, including the ones in Su and White (2007, 2008, 2014), the copular-based test in Bouezmarni, Rombouts, and Taamouti (2012), and the characteristic function-based test in Wang and Hong (2018).

Flaxman, Neill, and Smola (2015) highlights the advantage of incorporating non-IID observations in causal inference. The authors state that if spatial and temporal variation is properly accounted for, the resulting causal graphs could be more reasonable. A regression-based CI test is proposed in (Flaxman et al., 2015) to deal with non-IID data.

4 EMPIRICAL COMPARISON

We next examine the empirical performances of some representative CI tests, including KCIT by K. Zhang et al. (2012), CDIT by Wang et al. (2015), RCIT and RCoT by Strobl et al. (2019), CCIT by Sen et al. (2017), KRESIT by H. Zhang et al. (2017), CMIknn by Runge (2018), and GCM by Shah and Peters (2019). Codes of all these tests are freely available online.¹ For CMIknn, the option “transform = ‘ranks’” is used as in Runge (2018). CCIT is run with 30 bootstrap iterations in

FIGURE 1 Test performance comparisons with post-nonlinear noise model and similar setup as in Strobl et al. (2019). Shown are KS (left column), AUPC (center column), and runtime (right column) for an experiment with \( d_Z = 1 \) (top row), \( d_Z = 5 \) (middle row), and \( d_Z = 10 \) (bottom row). The original CPU time \( t \) is transformed in this way, \( \log(t + 1) \), for calculating runtime. The horizontal axis corresponds to the sample size. Error bars give the bootstrapped standard errors. AUPC, area under the power curve; KS, Kolmogorov–Smirnov
compliance with the example in its readme file. As the default setting of KRESIT is too time consuming, we adjust the parameters to “num_shuffles=300, max_iter=2000, learning_rate=0.05.” For other packages, we use their default settings.

The synthetic data are generated according to the post-nonlinear noise model described in K. Zhang et al. (2012) and Strobl et al. (2019). For comparing type-I error rates, the data generating procedure goes as follows: \( X = g_1 \left( \frac{1}{dZ} \sum_{i=1}^{dZ} Z_i + \varepsilon_1 \right) \), \( Y = g_2 \left( \frac{1}{dZ} \sum_{i=1}^{dZ} Z_i + \varepsilon_2 \right) \), where \( Z = (Z_1, Z_2, \ldots, Z_{dZ}) \), \( \varepsilon_1 \) and \( \varepsilon_2 \) have jointly independent standard Gaussian distributions, \( g_1 \) and \( g_2 \) are smooth functions chosen uniformly from the set \{\( \cdot \), \( \cdot \)^2, \( \cdot \)^3, \( \tanh(\cdot) \), \( \exp(\| \|_2) \} \). Therefore, the data satisfies \( X \perp Y \mid Z \).

To evaluate the statistical power, we add the same small error \( \varepsilon_b \sim \mathcal{N}(0, 0.8^2) \) to both \( X \) and \( Y \) as follows: \( X = g_1(\varepsilon_b + \varepsilon_1) \), \( Y = g_2(\varepsilon_b + \varepsilon_2) \), \( Z \sim \mathcal{N}(0, I_{dZ}) \), where \( I_{dZ} \) is a \( dZ \)-dimensional identity matrix. In this case, we have \( X \not\perp Y \mid Z \). All data are normalized.

The comparisons are done for sample sizes 50, 100, 200, 400, 700, and 1,000, the conditional dimension \( dZ \) taking values in \{1, 5, 10\}. As in Strobl et al. (2019), we calculate the Kolmogorov–Smirnov (KS) statistic as a metric to quantify how uniform the distribution of \( p \)-values is, under the null hypothesis. A smaller KS value indicates a better test. We evaluate the test power by computing the area under the power curve (AUPC), which is equivalent to the area under the empirical CDF of the \( p \)-values returned by a certain CI test when the null does not hold. Thus, tests with AUPC closer to 1 have higher power.

The simulation results are shown in Figure 1. All metrics were evaluated from 100 realizations of the model and error bars give the bootstrapped standard errors. KCIT performs well in most of the cases except for small sample sizes. RCIT and RCoT tend to give similar results, but RCoT outperforms RCIT as \( dZ \) increases. As RCIT and RCoT are approximation based, they are unreliable for sample sizes under 100. CDIT seems not applicable for high-dimensional cases. In Wang et al. (2015), simulations are conducted only for \( dZ \leq 2 \). In our experiment, CDIT has AUPC below 0.1 for all sample sizes when \( dZ = 10 \), so the CDIT curve is not shown in Figure 1h. The performance of CCIT might be improved by increasing the bootstrap iteration number. And since CCIT involves splitting input dataset into two groups for constructing data of two classes (after which one can choose a classification method for performing CI test), it has low AUPC for small sample sizes. The strength of CCIT lies on scalability to high-dimensional data with \( dZ > 20 \), which is out of the scope of our experiments. CMIknn and GCM are comparably more stable in the sense that they are well calibrated for almost all circumstances. KRESIT is found to be very time consuming; it takes nearly 2 hr to test one dataset of sample size 1,000, we therefore only provide simulation results of KRESIT for sample sizes 50, 100, 200, and 400. In these experiments, KRESIT exhibits impressively high performance.

### TABLE 1 MCC comparisons with post-nonlinear noise model and similar setup as in Strobl et al. (2019)

| \( dZ \) | CDIT: 0.3477 | CMIknn: 0.3727 | GCM: 0.5263 | KCIT: 0.5674 | KRESIT: 0.5799 | RCoT: 0.7798 |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | | | | | | |
| 5 | CDIT: 0.3384 | CMIknn: 0.342 | KRESIT: 0.4973 | KRESIT: 0.5658 | RCoT: 0.6499 | RCoT: 0.7294 |
| 10 | CMIknn: 0.1846 | CMIknn: 0.325 | CMIknn: 0.3828 | KCIT: 0.4201 | KCIT: 0.4733 | KCIT: 0.61 |
| Total | CMIknn: 0.2064 | CMIknn: 0.317 | KRESIT: 0.4503 | KRESIT: 0.53 | RCoT: 0.5967 | RCoT: 0.6507 |

Note: Shown are top three MCCs for different settings (\( \alpha = .05 \)), for example, cell in Row 2 and Column 2 gives the top three test methods ranked by their MCCs, for the case \( dZ = 1 \) and sample size equals 50; cell in the last row and the last column gives the top three test methods ranked by their MCCs, which is calculated by pooling together \( p \)-values of cases with sample size 1,000 and all \( dZ \)’s.

Abbreviations: MCC, Matthews correlation coefficient.
Tests with low KS statistics may also have low AUPCs, and tests with high AUPCs may get high KS statistics. For a more comprehensive criterion, Table 1 lists the top three Matthews correlation coefficients (MCCs) of different settings (with significant level $\alpha = .05$). Since a main application of CI tests is graph model structure learning, where a selected CI test would be used for testing data with different $d_J$'s at the same time, it is reasonable for us to pay more attention on the bottom row of Table 1, where MCCs are calculated by pooling together $p$-values of a given sample size and all $d_J$'s. The results indicate that, for $n$ is around 50, CMIknn and GCM outperform the others. For $n$ range from 100 to 400, KRESIT is not only the top 1 test in the last row, it is top 1 or is very close to top 1 for all other cases. But if runtime is of prime concern, CMIknn and RCoT are more preferred. For data of larger sizes, such as $n$ equals 700 or 1,000, RCoT could be the best choice. Note that KRESIT is not compared in these cases because of its computational burden, yet we can expect it to have high performances. After all, as mentioned in Section 3.1.5, KRESIT and RCoT are essentially estimating the same test statistic, and RCoT can be viewed as a large-scale approximation of KRESIT (H. Zhang et al., 2017).

5 | CONCLUSION

Since no nontrivial CI test can control the Type-I error for the entire class of distributions satisfying CI, the users of nonparametric CI tests need both the understanding of available CI tests and the knowledge of the problem in hand in order to choose the right test. As an effort to service this need, we provide this brief review of nonparametric CI tests. The review begins with a brief introduction to classical applications and the development of CI tests, then provides a short review of the theoretical background of CI, and discusses different angles for constructing nonparametric CI tests. In the end, we construct a simulation study comparing the performance of several popular tests.

From the empirical comparison, we see that CI testing is still an open challenge. As demanded by the growing interest on probability graphical models and casual inference, CI tests with lower computational complexity and the ability to solve higher dimensional problems attract researchers’ interests.

Recently, some new ideas emerge. Burkart and Király (2017) address CI testing problems by establishing a theoretical link between CI testing and supervised learning task, which enables extensively studied supervised learning workflows to be directly transferred to independence testing workflows. Sen, Shanmugam, Asnani, Rahimzamani, and Kannan (2018) cast this problem under a so-called meta-algorithm, “Mimic and Classify,” which (a) mimics the CI distribution close enough to recover the support, and (b) classifies to distinguish the joint and the CI distribution. By doing so, CI testing can be handled by methods from the modern advances in Deep Learning. For further reading on CI testing and other domains, we direct the reader to Pearl (1998), Koller and Friedman (2009), Scutari (2010), and Murphy (2012).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Chun Li: Conceptualization; formal analysis; funding acquisition; investigation; methodology; software; validation; visualization; writing. Xiaodan Fan: Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing; writing, review, and editing

ORCID

Xiaodan Fan © https://orcid.org/0000-0002-2744-9030

ENDNOTE

1 KCIT, RCIT, and RCoT: https://github.com/ericstrobl/RCIT. CDIT: R package cdcsis. GCM test: R package GeneralisedCovarianceMeasure. CCIT: https://github.com/rajatsen91/CCIT. KRESIT: https://github.com/oxmlcs/kery. CMIknn: https://github.com/jakobrunge/tigramite.
ORCID

Xiaodan Fan https://orcid.org/0000-0002-2744-9030

RELATED WIRES ARTICLES

Kernel-based measures of association

REFERENCES

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In J. van den Bussche & V. Vianu (Eds.), Database theory—ICDT 2001 (Vol. 1973, pp. 420–434). Berlin, Heidelberg: Springer.

Bergsma, W. P. (2004). Testing conditional independence for continuous random variables. *Eurandom Report, 049.* https://www.eurandom.tue.nl/reports/2004/048-report.pdf

Bergsma, W. P. (2010). *Nonparametric Testing of Conditional Independence by Means of the Partial Copula* (SSRN Scholarly Paper No. ID 1702981). Retrieved from https://papers.ssrn.com/abstract=1702981.

Blitzstein, J. K. (2011). *Problems in probability: Quotes.* Retrieved from http://www.hcs.harvard.edu/cs50-probability/quotes.php.

Bouezmarni, T., Rombouts, J. V. K., & Taamouti, A. (2012). Nonparametric copula-based test for conditional independence with applications to Granger causality. *Journal of Business & Economic Statistics, 30*(2), 275–287. https://doi.org/10.1080/07350015.2011.638831

Burkart, S., & Király, F. J. (2017). Predictive independence testing, predictive conditional Independence testing, and predictive graphical modelling. *ArXiv:1711.05869 [Cs, Math, Stat].* Retrieved from http://arxiv.org/abs/1711.05869.

Dawid, A. J. (1980). Partial association measures and an application to qualitative regression. *Biometrika, 67*(3), 581–590. https://doi.org/10.1093/biomet/67.3.581

Dawid, A. P. (1979). Conditional Independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological), 41*(1), 1–15. https://doi.org/10.1111/j.2517-6161.1979.tb01052.x

Dawid, A. P. (1980). Conditional independence for statistical operations. *The Annals of Statistics, 8*(3), 598–617.

Doran, G., Muandet, K., Zhang, K., & Scholkopf, B. (2014). A permutation-based kernel conditional Independence test. Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (pp. 132–141).

Edwards, D. (2000). *Introduction to graphical modelling.* New York, NY: Springer Science & Business Media.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507. https://doi.org/10.2307/2331838

Fisher, R. A. (1921). On the probable error of a coe cient of correlation deduced from a small sample. *Metron, 1*, 3–32.

Flaxman, S. R., Neill, D. B., & Smola, A. J. (2015). Gaussian processes for independence tests with non-iid data in causal inference. *ACM Transactions on Intelligent Systems and Technology, 7*(2), 1–23. https://doi.org/10.1145/2806892

Florens, J.-P., & Fougere, D. (1996). Noncausality in continuous time. *Econometrica, 64*(5), 1195. https://doi.org/10.1207/2271962

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems, 20*, 489–496.

Geiger, D., & Heckerman, D. (1994). Learning Gaussian networks. Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence (pp. 235–243). Retrieved from http://dl.acm.org/citation.cfm?id=2074394.2074425.

Gerstein, M. B., Kundaje, A., Harinaran, M., Landt, S. G., Yan, K.-K., Cheng, C., … Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature, 489*, 91–100.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424. https://doi.org/10.1207/2271962

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research, 13*, 723–773.

Huang, T. M., & Aggarwal, C. C. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics, 38*(4), 2047–2091. https://doi.org/10.1214/09-AOS770
