Statistica Sinica Preprint No: SS-2026-0108	
Title	A Family of Chatterjee's Correlation Coefficients and Their Properties
Manuscript ID	SS-2026-0108
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0108
Complete List of Authors	Muhong Gao and Qizhai Li
Corresponding Authors	Qizhai Li
E-mails	liqz@amss.ac.cn
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

A FAMILY OF CHATTERJEE'S CORRELATION COEFFICIENTS AND THEIR PROPERTIES

Muhong Gao and Qizhai Li*

University of International Business and Economics,

and Academy of Mathematics and System Science, Chinese Academy of Sciences

Abstract: Quantifying the strength of functional dependence between random scalars X and Y is a fundamental statistical problem. While many existing correlation coefficients effectively identify linear or monotone functional dependence, they often fail to capture more general, non-monotonic dependencies. In response, we propose a family of simple and computationally efficient correlation coefficients that encompasses a wide class of novel dependency measures, including the Chatterjee's correlation coefficient as a special case. We prove that each of our correlation coefficients converges to a population-level correlation measure as the sample size grows, and a central limit theorem is established under the assumption of independence between X and Y. Moreover, we provide the necessary and sufficient condition for the correlation measure to satisfy three desirable properties: (P1). it belongs to the range of $[0, 1]$; (P2). it equals 1 if and only if Y is a measurable function of X; and (P3). it equals 0 if and only if Y is independent of X. Based on both theoretical analysis and empirical evi-

*Corresponding author

Statistica Sinica: Newly accepted Paper

dence, our correlation coefficients demonstrate superior practicality, robustness, and computational efficiency compared with others.

Key words and phrases: Central limit theorem, correlation coefficient, functional dependence, independence, rank-based correlation.

1. Introduction

Measuring the strength of functional dependence between two random scalars X and Y is a fundamental problem in statistics with broad scientific implications. In explicit statistical terms, the objective is to formulate a straightforward dependence estimate $\xi(X, Y)$, wherein a higher value of $\xi(X, Y)$ indicates a stronger dependence of Y on X. Ideally, $\xi(X, Y)$ attains its maximum value if and only if $Y = f(X)$, where f represents an arbitrary non-random measurable function. From an information-theoretic perspective, quantifying functional dependence is also closely related to the core problem of measuring the information gain about Y from observing X (Tishby et al., 1999) or the shared randomness between X and Y (Sudan et al., 2020). Across the extensive literature spanning the past century that focuses on quantifying pairwise associations, the majority of proposals have been primarily crafted for identifying linear/monotonic dependence patterns or testing independence, including the classical Pearson's corre-

Statistica Sinica: Newly accepted Paper

lation coefficient, Spearman's ρ, and Kendall's τ, as well as the distance correlations (Székely et al., 2007; Székely and Rizzo, 2009; Wang et al., 2015), rank correlations (Drton et al., 2020; Weihs et al., 2018; Bergsma and Dassios, 2014; Shi et al., 2022), copula-based correlations (Schweizer and Wolff, 1981; Chang et al., 2016; Lopez-Paz et al., 2013; Zhang, 2019), binning-based correlations (Heller et al., 2012; Kinney and Atwal, 2014; Ma and Mao, 2019), among many others. While for measuring the amount of general functional dependence where f is possibly to be non-monotone, these approaches typically fall short in delivering meaningful and efficacious assessments.

1.1 Chatterjee's correlation coefficient

The recent seminal work of Chatterjee (2021) has attracted very much attention. It proposed a new elegant correlation coefficient that effectively measures the extent of functional dependence under arbitrary f functions. Given the i.i.d. data copies $(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$ with $n \geq 2$, the Chatterjee's correlation coefficient has a very simple expression:

$$\xi_n(X, Y) = 1 - \frac{n^{-2} \cdot \sum_{i=1}^{n-1} |R(Y_{[i+1]}) - R(Y_{[i]})|}{\chi_n(Y)}, \tag{1.1}$$

Statistica Sinica: Newly accepted Paper

where $(X_{(1)}, Y_{[1]}), \ldots, (X_{(n)}, Y_{[n]})$ are reordered pairs of data such that $X_{(1)} \leq \cdots \leq X_{(n)}$, and $R(Y_{[i]}) = \sum_{j=1}^{n} \mathrm{I}(Y_j \leq Y_{[i]})$ represents the rank of $Y_{[i]}$ among all Y_j's. The denominator term $\chi_n(Y)$, acting as a normalization statistic, equals $(n^2-1)/(3n^2)$ if there are no ties among X_i's, and has a slightly altered expression in the presence of ties among X_i's. Despite its simple nonparametric form, the most compelling feature of Chatterjee's correlation coefficient is its almost-surely convergence to a limit $\xi = \lim_{n\to\infty} \xi_n(X, Y)$, which satisfies the following three desirable properties:

(P1) (Normalization). $\xi \in [0, 1]$, for all distributions of (X, Y).

(P2) (Perfect dependence). $\xi = 1$ (reaches its maximum) if and only if there is a measurable function $f : \mathbb{R} \to \mathbb{R}$, such that $Y = f(X)$ almost surely.

(P3) (D-Consistency). $\xi = 0$ (reaches its minimum) if and only if X and Y are independent.

These three properties of ξ, as first introduced by a prior work (Dette et al., 2013), appear to be so strong that none of the preceding correlation measures can simultaneously fulfill all of them. In particular, the perfect dependence property (P2) stands as a distinguished characteristic of Chatterjee's correlation coefficient, granting $\xi_n(X, Y)$ extraordinary capability

Statistica Sinica: Newly accepted Paper

in reflecting functional dependence. Several other correlation coefficients share the property of being maximized when $Y = f(X)$ is satisfied, including the maximal correlation coefficient (Rényi, 1959), the maximum information coefficient (Reshef et al., 2011), and the Hellinger correlation (Geenens and Lafaye de Micheaux, 2022). But the converse is not true: they may also achieve maximum values when Y is not a function of X. For example, the maximum information coefficient can reach its maximum value when $Y^2 = X$, in which case Y cannot be represented as a function of X. Consequently, their identification of functional dependence can sometimes be inaccurate or misleading.

1.2 Motivation

We first present an intuitive rationale for the efficacy of Chatterjee's correlation coefficient (1.1). One ingenious aspect of (1.1) resides in the formulation of the numerator term $n^{-2}\sum_{i=1}^{n-1}|R(Y_{[i+1]}) - R(Y_{[i]})|$. Here, $|R(Y_{[i+1]}) - R(Y_{[i]})|$ is noted for quantifying the degree of variation between two consecutive $Y_{[i]}$'s. Considering that the $(X_{(i)}, Y_{[i]})$ pairs are rearranged based on the order of $X_{(i)}$ values, the variation between two consecutive $Y_{[i]}$ and $Y_{[i+1]}$ tends to be small when the dependence of Y on X is significant. Conversely, a greater variation is likely to occur when Y exhibits greater

Statistica Sinica: Newly accepted Paper

independence from X. Therefore, achieving the maximum or minimum of (1.1) fundamentally relies on the dependence level between Y and X.

This motivates a natural conjecture regarding potential extensions of Chatterjee's correlation coefficient: replacing the term $|R(Y_{[i+1]}) - R(Y_{[i]})|$ in (1.1) by other types of metrics on the degree of variation between $Y_{[i]}$ and $Y_{[i+1]}$ may still be effective for measuring functional dependence. This paper is dedicated to conducting a detailed investigation into this interesting question.

1.3 Our contributions

We propose a new correlation coefficient

$$\xi_n^{(h,F)}(X,Y) = 1 - \frac{n^{-1} \cdot \sum_{i=1}^{n-1} h\big\{F(Y_{[i]}), F(Y_{[i+1]})\big\}}{\chi_n^{(h,F)}(Y)}, \tag{1.2}$$

with $F(\cdot)$ being a monotone univariate function, and $h(\cdot,\cdot)$ being a non-negative continuous bivariate function, satisfying $h(x,x) = 0$ for $x \in \mathbb{R}$. To enhance identifiability, we further restrict $F(\cdot)$ to be a cumulative distribution function (c.d.f.). The denominator $\chi_n^{(h,F)}(Y)$ serves as a normalization statistic, whose explicit expression will be presented later in Section 2. Comparing (1.2) with (1.1), the term $h\{F(Y_{[i]}), F(Y_{[i+1]})\}$ in (1.2) is an

Statistica Sinica: Newly accepted Paper

extension of the corresponding term $n^{-1} \cdot |R(Y_{[i+1]}) - R(Y_{[i]})|$ in (1.1), capable for quantifying the degree of variation between $Y_{[i]}$ and $Y_{[i+1]}$. With a variety of choices for h and F functions, $\xi_n^{(h,F)}(X,Y)$ encompasses a wide family of new correlation coefficients that have not been previously explored in existing literature. It is noticeable that the Chatterjee's correlation coefficient (1.1) corresponds to a special case of $\xi_n^{(h,F)}(X,Y)$ in (1.2) when taking $h(x,y) = |x-y|$, and $F(x) = F_{Y,n}(x) = 1/n \sum_{i=1}^{n} \mathrm{I}(Y_i \leq x)$ as the empirical c.d.f. of Y_i's.

This paper seeks to provide answers to the fundamental questions concerning the asymptotic behaviors and limit properties of $\xi_n^{(h,F)}(X,Y)$, which can be contingent upon the selection of h and F. We obtain a series of valuable results as outlined below.

1. $\xi_n^{(h,F)}(X,Y)$ converges almost surely to a limit $\xi^{(h,F)}$.

2. The limit $\xi^{(h,F)}$ satisfies the perfect dependence property (P2) and the sufficiency part of property (P3) for almost all continuous functions h and F, under only mild regularity assumptions.

3. The limit $\xi^{(h,F)} = 0$ as long as X and Y are independent. Moreover, a central limit theorem is established for $\xi_n^{(h,F)}(X,Y)$ in the case where X and Y are independent.

Statistica Sinica: Newly accepted Paper

4. We establish the *necessary and sufficient* condition for $\xi^{(h,F)}$ to satisfy all properties (P1)–(P3). Specifically, we delineate the set $\mathcal{H}$ comprising continuous bivariate functions, wherein $\xi^{(h,F)}$ possesses properties (P1)–(P3) if and only if $h \in \mathcal{H}$. Our findings reveals a one-to-one correspondence between $\mathcal{H}$ and the family of bivariate characteristic kernel functions, suggesting that $\mathcal{H}$ is as rich as the latter.

5. While the coefficient $\xi_n^{(h,F)}(X,Y)$ is initially formulated using deterministic h and F functions, we also allow taking $F = F_{Y,n}$ as the empirical c.d.f. of Y_i's. In this scenario, $\xi_n^{(h,F_{Y,n})}(X,Y)$ becomes a rank-based correlation coefficient, converges to $\xi^{(h,F_Y)}$, and retains the majority of the four results described above. The only difference is that we establish a sufficient condition for $\xi^{(h,F_Y)}$ to satisfy properties (P1)–(P3), while establishing the corresponding necessary condition remains an open problem.

The above results admit the existence of a wide family of new correlation coefficients that have theoretically-backed capability in measuring functional dependence between bivariate random variables. When dealing with real-world problems, our $\xi_n^{(h,F)}(X,Y)$ can better cater to practical needs by adapting h and F functions in a manner tailored to specific purposes. This flexibility contrasts with the constrained structure of the Chatterjee's corre-

Statistica Sinica: Newly accepted Paper

lation coefficient. Extensive numerical experiments lend empirical support to the practical efficacy of our proposals.

Outline of the paper. Section 2 formally presents our correlation coefficient $\xi_n^{(h,F)}(X,Y)$ alongside its fundamental asymptotic properties. Section 3 demonstrates the necessary and sufficient condition for achieving all desired properties (P1)–(P3). Section 4 investigates the properties of the rank-based variant of $\xi_n^{(h,F)}(X,Y)$ when taking $F = F_{Y,n}$. Section 5 and 6 conduct numerical experiments on both synthetic data and real-world data. Section 7 briefly concludes. The Supplementary Material contains the proofs of main results.

Notations. For integer $n \geq 2$, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples of (X, Y) in the probability space $\{\Omega, \mathscr{F}, \mathrm{P}\}$. Denote by $(X_{(1)}, Y_{[1]}), \ldots, (X_{(n)}, Y_{[n]})$ the reordered pairs of samples such that $X_{(1)} \leq \cdots \leq X_{(n)}$. If there are ties among X_i's, breaking the ties uniformly at random in this ordering. Let $\mathrm{I}(\cdot)$ denote the indicator operator. Let $F_X(t) := \mathrm{P}(X \leq t)$ and $F_Y(t) := \mathrm{P}(Y \leq t)$ be the c.d.f.'s of X and Y respectively, and let $F_{Y,n}(t) = 1/n \sum_{i=1}^n \mathrm{I}(Y_i \leq t)$ be the empirical c.d.f. of Y_i's. Denote by $\Phi(\cdot)$ the c.d.f of standard normal distribution. Denote by μ_X and μ_Y the distribution laws of X and Y. Let $\mu_{Y|X=x}$ (or $\mu_{Y|x}$ for short) denote the

Statistica Sinica: Newly accepted Paper

conditional law of Y given $X = x$. For the existence of $\mu_{Y|x}$, refer to Theorem 2.1.22 and Exercise 4.1.18 in Durrett (2019) (also see Lemma 1 in the Supplementary Material). Let $\xrightarrow{\mathrm{P}}$ and $\xrightarrow{\mathcal{D}}$ denote convergence in probability and in distribution respectively. For simplicity, the term "almost surely" is abbreviated as "a.s." in case needed.

2. The new family of correlation coefficients

In this section, we formally present the formulation of our proposed correlation coefficient along with its fundamental properties. Throughout the paper, we assume that neither X nor Y is a constant.

Definition 1 (New correlation coefficient)**.** For a non-negative continuous bivariate function $h : [0,1]^2 \to [0,\infty)$ satisfying $h(x,x) = 0$ for all $x \in [0,1]$, and a c.d.f. function $F : \mathbb{R} \to [0,1]$, define the variation statistic

$$\zeta_n^{(h,F)}(X,Y) := \frac{1}{n}\sum_{i=1}^{n-1} h\big\{F(Y_{[i]}), F(Y_{[i+1]})\big\},$$

and the normalization statistic

$$\chi_n^{(h,F)}(Y) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h\{F(Y_i), F(Y_j)\}. \tag{2.1}$$

Statistica Sinica: Newly accepted Paper

The new correlation coefficient is defined as:

$$\begin{aligned}\xi_n^{(h,F)}(X,Y) &:= 1-\zeta_n^{(h,F)}(X,Y)/\chi_n^{(h,F)}(Y)\\ &= 1-\frac{n\sum_{i=1}^{n-1}h\big\{F(Y_{[i]}),F(Y_{[i+1]})\big\}}{\sum_{i=1}^{n}\sum_{j=1}^{n}h\{F(Y_i),F(Y_j)\}}.\end{aligned}$$

In case that $\chi_n^{(h,F)}(Y)=0$, we set $\xi_n^{(h,F)}(X,Y)=1$.

As previously mentioned in Section 1.3, the numerator term $\zeta_n^{(h,F)}(X,Y)$ is designed to measure the degree of variation between consecutive $Y_{[i]}$'s. The denominator term $\chi_n^{(h,F)}(Y)$ serves as a normalization statistic, ensuring that $\xi_n^{(h,F)}(X,Y)$ tends towards zero if X and Y are independent. Theorem 1 shows that $\xi_n^{(h,F)}(X,Y)$ converges almost surely to a deterministic constant $\xi^{(h,F)}$ as n approaches infinity.

Theorem 1 (Convergence of correlation coefficient)**.** *As $n\to\infty$, we have the following three convergence results:*

(i) *the variation statistic $\zeta_n^{(h,F)}(X,Y)$ converges a.s. to the limit*

$$\zeta^{(h,F)} := \int\int\int h\big\{F(y),F(z)\big\}\,\mathrm{d}\mu_{Y|x}(y)\,\mathrm{d}\mu_{Y|x}(z)\,\mathrm{d}\mu_X(x);$$

Statistica Sinica: Newly accepted Paper

(ii) *the normalization statistic* $\chi_n^{(h,F)}(Y)$ *converges a.s. to the limit*

$$\chi^{(h,F)} := \int\int h\{F(y), F(z)\}\ \mathrm{d}\mu_Y(y)\ \mathrm{d}\mu_Y(z);$$

and (iii) *the correlation coefficient* $\xi_n^{(h,F)}(X,Y)$ *converges a.s. to the limit*

$$\xi^{(h,F)} \ := \ 1 - \zeta^{(h,F)}/\chi^{(h,F)}. \tag{2.2}$$

In case that $\chi^{(h,F)} = 0$*, we set* $\xi^{(h,F)} = 1$*.*

Throughout the paper, we call the population-level deterministic quantity $\xi^{(h,F)}$ the "correlation measure", to distinguish it from its consistent empirical estimator $\xi_n^{(h,F)}(X,Y)$ which is called the "correlation coefficient". Several remarks are provided in order.

Remark 2.1. The new correlation coefficient $\xi_n^{(h,F)}(X,Y)$ requires no need for estimating probability densities, characteristic functions or mutual information, and it could be simply computed with a minimum time complexity of $O(n \log n)$ (see Remark 2.9 for detail).

Remark 2.2. In Theorem 1, there are no constraints on the distribution of (X,Y) except that neither X nor Y is a constant. Thus Theorem 1 is applicable to both discrete and continuous (X,Y) distributions.

Statistica Sinica: Newly accepted Paper

Remark 2.3. Theorem 1 is almost assumption-free on h and F, except that h is required to be continuous (as specified in Definition 1). This continuity condition for h is vital for establishing Theorem 1. Lemma A.6 in Section S1 of the Supplementary Material shows a trivial example of a discontinuous $h(y,z) = \mathrm{I}(y \neq z)$ for which $\xi_n^{(h,F)}(X,Y)$ does not converge to $\xi^{(h,F)}$.

Remark 2.4. When taking $h(y,z) = |y-z|$ and $F = F_Y$ as the c.d.f. of Y in (2.2), it could be verified (see Lemma A.7 in Section S1 of the Supplementary Material) that the resulting correlation measure $\xi^{(h,F_Y)}$ has an alternative expression

$$\xi^{(h,F_Y)} = 1 - \frac{\zeta^{(h,F_Y)}}{\chi^{(h,F_Y)}} = \frac{\int \mathrm{var}\big\{\mathrm{E}(\,\mathrm{I}(Y \geq t) \mid X)\big\}\,\mathrm{d}\mu_Y(t)}{\int \mathrm{var}\big\{\,\mathrm{I}(Y \geq t)\big\}\,\mathrm{d}\mu_Y(t)},$$

which coincides with the correlation measure ξ in Chatterjee (2021) [Theorem 1.1, Eq. (2)].

Remark 2.5. Our correlation coefficients are generally not symmetric in X and Y. This asymmetry, inherited from Chatterjee's correlation coefficient, arises because the underlying task of measuring functional dependence is itself asymmetric: the fact that Y is a function of X does not, in general, imply that X is also a function of Y. If a symmetric version of the corre-

Statistica Sinica: Newly accepted Paper

lation coefficient is desired, one may consider simple symmetrizations such as $\max\{\xi_n^{(h,F)}(X,Y), \xi_n^{(h,F)}(Y,X)\}$ or $\{\xi_n^{(h,F)}(X,Y) + \xi_n^{(h,F)}(Y,X)\}/2$. Also see Remark 1 of Chatterjee (2021) for further discussion on symmetry.

Remark 2.6. From our definition, it is noted that $\xi_n^{(h,F)}(X,Y)$ or $\xi^{(h,F)}$ fully depends on the functions h and F, yet this relationship is not "identifiable". In other words, for distinct pairs of functions (h_1, F_1) and (h_2, F_2), $\xi_n^{(h_1,F_1)}(X,Y)$ and $\xi_n^{(h_2,F_2)}(X,Y)$ could be identical correlation coefficients, provided that the composite functions $h_1\{F_1(\cdot), F_1(\cdot)\}$ and $h_2\{F_2(\cdot), F_2(\cdot)\}$ are the same. However, treating h and F separately is intentional, because h and F have different influences on the properties of the correlation coefficient. Such design of separate h and F not only facilitates further theoretical derivations, but also provides more concrete guidelines for constructing the correlation coefficients in real practice.

Remark 2.7. Another advantage of using the c.d.f. F in our correlation coefficient is that, it transforms the unbounded data Y into $F(Y)$ with a bounded range of $[0, 1]$, which makes the correlation coefficient more robust against the outliers of Y. Moreover, if F is removed from our correlation coefficient (or equivalently, taking F as the identity mapping), then the convergence in Theorem 1 may no longer hold for all continuous bivariate functions h.

Statistica Sinica: Newly accepted Paper

Remark 2.8. Note that the function h in $\xi_n^{(h,F)}(X,Y)$ is not required to be symmetric, which provides more flexibility. While for non-symmetric h, the correlation measure $\xi^{(h,F)}$ in (2.2) remains unchanged if replacing this h with the symmetric $h^\dagger(y,z) := \{h(y,z)+h(z,y)\}/2$. Thus, when it turns to study the properties of $\xi^{(h,F)}$, assuming h to be symmetric will not lose generality. For convenience, we assume that h is symmetric in the remainder of paper (refer to condition (C1)).

Remark 2.9. The computational time complexities of $\zeta_n^{(h,F)}(X,Y)$ and $\chi_n^{(h,F)}(Y)$ are in $O(n\log n)$ and $O(n^2)$ respectively, which leads to an overall time complexity of $O(n^2)$ for $\xi_n^{(h,F)}(X,Y)$. In fact, replacing $\xi_n^{(h,F)}(X,Y)$ by $\xi_m^{(h,F)}(X,Y)$ with $m = O(\sqrt{n\log n})$ does not affect the convergence results stated in Theorem 1, while reducing the computational complexity of to $O(n\log n)$. Therefore, a practical recommendation for improving computational efficiency is to use $\xi_m^{(h,F)}(X,Y)$ with $m = O(\sqrt{n\log n})$ in place of $\xi_n^{(h,F)}(X,Y)$ in (2.1). It is worth noting that this reduction in computational cost comes at the expense of an increased variance of the estimator, from $O(n^{-1})$ to $O((n\log n)^{-1/2})$.

In Definition 1, both h and F in $\xi_n^{(h,F)}(X,Y)$ are, in principle, prespecified deterministic functions. More generally, however, F can also be allowed to be a random c.d.f. estimated from the sample. One important special

Statistica Sinica: Newly accepted Paper

case is $F = F_{Y,n}$, the empirical c.d.f. of Y, which we discuss separately and in detail in Section 4. Another possibility is a general sample-based random c.d.f. $\widehat{F}_n$. Corollary 1 shows that, under certain conditions, the resulting correlation coefficient $\xi_n^{(h,\widehat{F}_n)}(X,Y)$ with $F = \widehat{F}_n$ plugged-in also enjoys almost-surely convergence.

Corollary 1 (Convergence of correlation coefficient with random $\widehat{F}_n$)**.** *Let F be a deterministic c.d.f.. Let $\{\widehat{F}_n\}_{n=1}^{\infty}$ be a sequence of random c.d.f.s, such that $\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \to 0$, a.s., as $n \to \infty$. Assume that h is Lipschitz continuous in $[0,1]^2$. Then as $n \to \infty$, the correlation coefficient $\xi_n^{(h,\widehat{F}_n)}(X,Y)$ converges to the limit $\xi^{(h,F)}$ a.s..*

In the real-data experiment in Section 6, we introduce a specific example involving the random c.d.f. $\widehat{F}_n(x) = \Phi((x - \widehat{\mu}_Y)/\widehat{\sigma}_Y)$ where $\widehat{\mu}_Y$ and $\widehat{\sigma}_Y$ are the sample mean and standard deviation of Y, respectively, and Φ denotes the c.d.f. of the standard normal distribution. Such a sample-adapted c.d.f., which centers and scales the distribution based on the observed data, often leads to more stable and satisfactory performance in practical applications.

2.1 Perfect dependence property

For $\xi^{(h,F)}$ to serve as a reliable measure for the strength of functional dependence, one of the most desirable properties is the perfect dependence

Statistica Sinica: Newly accepted Paper

property (P2). Theorem 2 verifies that $\xi^{(h,F)}$ indeed possesses property (P2), provided that h and F satisfy the regularity conditions below.

(C1) (Regularity condition on h). $h : [0,1]^2 \to [0,\infty)$ is a non-negative continuous symmetric function, satisfying $h(x,x) = 0$, and $h(x,y) > 0$ for all $x, y \in [0,1]$.

(C2) (Regularity condition on F). $F : \mathbb{R} \to (0,1)$ is a continuous and strictly-increasing c.d.f. function.

Theorem 2 (Properties of correlation measure)**.** *Assume that h and F satisfy the regularity conditions (C1) and (C2). Then the correlation measure $\xi^{(h,F)}$ in (2.2) possesses the perfect dependence property (P2). Moreover, $\xi^{(h,F)}$ also satisfies the I-consistency property (P3)′:*

(P3)′ *(I-Consistency). If X and Y are independent, then $\xi^{(h,F)} = 0$.*

For the above results, we give some remarks.

Remark 2.10. The sufficiency part of property (P2) in Theorem 2 can be illustrated as follows. When $Y = f(X)$, the conditional probability measure $\mu_{Y|x}$ reduces to the Dirac measure $\delta_{f(x)}$ at point $f(x)$. Thus we have

$$\zeta^{(h,F)} = \int\int\int h\{F(y), F(z)\}\, \mathrm{d}\delta_{f(x)}(y)\, \mathrm{d}\delta_{f(x)}(z)\, \mathrm{d}\mu_X(x)$$

Statistica Sinica: Newly accepted Paper

$$= \int h\big[F\{f(x)\}, F\{f(x)\}\big] \, \mathrm{d}\mu_X(x) = 0,$$

from which $\xi^{(h,F)} = 1$ follows.

Remark 2.11. The main purpose for imposing the regularity conditions (C1)and (C2), is to ensure that $h\{F(y), F(z)\}$ is non-zero for any distinct $y, z \in \mathbb{R}$. This further admits the necessity part of property (P2) in Theorem 2. If there exists distinct y, z such that $h\{F(y), F(z)\} = 0$, then any random variable Y with the support $\{y, z\}$ would lead to $\zeta^{(h,F)} = \chi^{(h,F)} = 0$ and $\xi^{(h,F)} = 1$, which violates property (P2). Note that condition (C2) is satisfied if F is a c.d.f. of some continuous random variable with support of $\mathbb{R}$.

Remark 2.12. The property (P3)′ in Theorem 2, which is called the "I-consistency" property (Weihs et al., 2018), corresponds to the sufficiency part of the "D-consistency" property (P3). Its proof can be seen as follows. If X and Y are independent, then the conditional probability measure $\mu_{Y|x}$ becomes μ_Y for all $x \in \mathbb{R}$. This gives that

$$\begin{aligned} \zeta^{(h,F)} &= \int\int\int h\big\{F(y), F(z)\big\} \, \mathrm{d}\mu_Y(y) \, \mathrm{d}\mu_Y(z) \, \mathrm{d}\mu_X(x) \\ &= \int\int h\big\{F(y), F(z)\big\} \, \mathrm{d}\mu_Y(y) \, \mathrm{d}\mu_Y(z) = \chi^{(h,F)}, \end{aligned}$$

Statistica Sinica: Newly accepted Paper

which yields $\xi^{(h,F)} = 1 - \zeta^{(h,F)}/\chi^{(h,F)} = 0$.

2.2 Central limit theorem

Measuring the amount of functional dependence is the major motivation for devising the correlation coefficient $\xi_n^{(h,F)}(X,Y)$, while testing independence is not our initial focus. In particular, the correlation measure $\xi^{(h,F)}$ under the mild regularity conditions (C1) and (C2) satisfies the I-consistency property (P3)′, whereas the stronger D-consistency property (P3) will be derived in the next Section 3, after imposing further restrictions on h. But still, we derive the central limit theorem for $\xi_n^{(h,F)}(X,Y)$ under independent X and Y, suggesting its potential utility for conducting independence tests if desired.

Theorem 3 (CLT of $\xi_n^{(h,F)}(X,Y)$ under independence)**.** *Assume that h and F satisfy the regularity conditions (C1) and (C2). Assume that X and Y are independent. Then as $n \to \infty$, we have*

$$\sqrt{n}\,\xi_n^{(h,F)}(X,Y) \xrightarrow{\mathcal{D}} N(0,\sigma^2),$$

Statistica Sinica: Newly accepted Paper

with

$$\sigma^2 = \frac{\mathrm{E}\big[h\{F(Y_1), F(Y_2)\}^2\big] - 2\,\mathrm{E}\big[h\{F(Y_1), F(Y_2)\}\,h\{F(Y_1), F(Y_3)\}\big] + \mathrm{E}\big[h\{F(Y_1), F(Y_2)\}\big]^2}{\mathrm{E}\big[h\{F(Y_1), F(Y_2)\}\big]^2}$$

being strictly-positive, where Y_1, Y_2, Y_3 *are* i.i.d. *copies of* Y.

The asymptotic variance σ^2 in Theorem 3 could be consistently estimated via constructing U-statistics on Y_i's; see details in Lemma A.13 in Section S1 of the Supplementary Material. In the special case of $h(y, z) = |y - z|$, $F = F_Y$, and Y being continuous, we have $\sigma^2 = 2/5$, which coincides with the variance τ^2 in Chatterjee (2021) [Theorem 2.2]. Also see Theorem 6 and Remark 4.1 in Section 4 for more related discussions.

Remark 2.13. The study of limiting theorems and independence testing for the original Chatterjee's correlation coefficient has developed rapidly in recent years. Among the important contributions, Shi et al. (2021) studies the power of the corresponding independence test against local alternatives and establishes the rate-suboptimality of the test based on Chatterjee's correlation coefficient. Auddy et al. (2024) further derives the exact detection threshold and limiting local power. Subsequently, Lin and Han (2022b) proposes a power-enhanced extension of the correlation coefficient that substantially improves the power of the corresponding independence test. For

Statistica Sinica: Newly accepted Paper

limiting theorems, Lin and Han (2022a) establishes a CLT for Chatterjee's correlation coefficient under general dependence between (X, Y), while Gao et al. (2026) and Chhaibi et al. (2026) derive explicit closed-form expressions for the asymptotic variance in this CLT. It is natural to expect that these results may extend to our broader family of correlation coefficients. In particular, under suitable conditions on (h, F), we anticipate that $\xi_n^{(h,F)}(X, Y)$ may admit analogous general CLTs and limiting local power results. Establishing such extensions would provide a valuable direction for future research.

3. Pursuing all properties (P1)–(P3)

Owing to the perfect dependence property (P2), the family of correlation coefficients $\{\xi_n^{(h,F)}(X, Y) : h \text{ and } F \text{ satisfies conditions (C1) and (C2)}\}$ are readily capable for measuring strength of dependence of Y on X. To explore further understandings for the correlation coefficient, a natural question arises: for which cases of h and F does the corresponding correlation measure $\xi^{(h,F)}$ satisfy all the three desirable properties (P1)–(P3)? This section unveils the answer to this question.

Statistica Sinica: Newly accepted Paper

3.1 Pursuing property (P1)

We first inspect the "normalization property" (P1). From the definition of correlation measure $\xi^{(h,F)} = 1 - \zeta^{(h,F)}/\chi^{(h,F)}$ in (2.2), achieving property (P1) means that the difference $\chi^{(h,F)} - \zeta^{(h,F)}$ is non-negative under any distributions of (X, Y). Proposition 1 shows that this is achieved if and only if an additional condition (C3) is imposed on function h.

(C3) (Equivalence condition on h for achieving (P1)). There exists a continuous positive-definite symmetric function $\varphi : [0,1]^2 \to \mathbb{R}$, such that

$$h(y,z) = 2^{-1}\{\varphi(y,y) + \varphi(z,z)\} - \varphi(y,z), \qquad \text{for } y, z \in [0,1]. \tag{3.1}$$

Here, we call a symmetric bivariate function φ "positive-definite", if for any real numbers $y_1, \ldots, y_m \in [0,1]$ and $\omega_1, \ldots, \omega_m \in \mathbb{R}$ with $m \geq 1$, $\sum_{i=1}^m \sum_{j=1}^m \omega_i \omega_j \varphi(y_i, y_j) \geq 0$ holds. We also call φ a "kernel function" if it is positive-definite. The link between h and φ in (3.1) is also known as the Schoenberg correspondence (Schoenberg, 1938), which was originally established to relate a conditionally negative definite function h with zero diagonal to a positive-definite kernel φ.

Statistica Sinica: Newly accepted Paper

Proposition 1 (Necessary and sufficient condition for property (P1))**.** *Assume that h and F satisfy the regularity conditions (C1) and (C2). Then the correlation measure $\xi^{(h,F)}$ satisfies property (P1) if and only if h satisfies condition (C3).*

Remark 3.1. We provide some sketch of proof for Proposition 1. A closer inspection on $\chi^{(h,F)}$ and $\zeta^{(h,F)}$ in Theorem 1 yields a useful identity: $\chi^{(h,F)} - \zeta^{(h,F)} =$

$$\int \left[\int \int -h\big\{F(y), F(z)\big\} \left\{ \mathrm{d}\mu_{Y|x}(y) - \mathrm{d}\mu_Y(y) \right\} \left\{ \mathrm{d}\mu_{Y|x}(z) - \mathrm{d}\mu_Y(z) \right\} \right] \mathrm{d}\mu_X(x) \quad (3.2)$$

Note that the double integral in the square bracket is symmetric with respect to y and z, with $-h\big(F(y), F(z)\big)$ being its integral kernel. Thus, any negative-definite function h would guarantee the positiveness of this double integral. This partly reflects why h in (3.1) should include a term "$-\varphi(y,z)$", which is negative-definite. After incorporating $-\varphi(y,z)$ into h, the purpose for adding the other two terms $\varphi(y,y)$ and $\varphi(z,z)$ is to ensure that h is nonnegative and $h(y,y) = 0$ is satisfied for all $y \in [0,1]$.

Remark 3.2. Condition (C3) and the link in (3.1) is also related to energy distance and kernel-based statistics. Note that the integrand $\int \int -h\big\{F(y), F(z)\big\} \left\{ \mathrm{d}\mu_{Y|x}(y) - \mathrm{d}\mu_Y(y) \right\} \left\{ \mathrm{d}\mu_{Y|x}(z) - \mathrm{d}\mu_Y(z) \right\}$ in (3.2) can be viewed as the general energy

Statistica Sinica: Newly accepted Paper

distance $D_{E,\rho}$ (Lyons, 2013) with metric $\rho = h$, which measures the discrepancy between the distributions $\mu_{Y|x}$ and μ_Y. Sejdinovic et al. (2013) showed that, for $D_{E,\rho}$ to be nonnegative, ρ needs to admit the Schoenberg correspondence in (3.1) with some positive-definite kernel function φ. Through this correspondence, $D_{E,\rho}$ can be expressed in terms of the maximum mean discrepancy (MMD) (Gretton et al., 2012), $\|m_\varphi(\mu_{Y|x}) - m_\varphi(\mu_Y)\|_{\mathcal{H}_\varphi}$, with kernel φ. This connection between the energy distance and MMD is established through the correspondence in (3.1), which is precisely the same structural condition that we identify in condition (C3) as the necessary and sufficient condition for our measure to be nonnegative.

3.2 Pursuing properties (P1)–(P3)

Combining the results in Theorem 2 and Proposition 1, we have derived the necessary and sufficient condition such that $\xi^{(h,F)}$ satisfies properties (P1), (P2), and (P3)′. As long as the normalization property (P1) and the "I-consistency" property (P3)′ are both achieved, $\xi^{(h,F)}$ is guaranteed to attain its minimum value of 0 under independent X and Y. To fulfill all the three desired properties (P1)–(P3), the last step is to upgrade property (P3)′ to property (P3). That is, we just need to ensure that there exists no dependent X and Y such that $\xi^{(h,F)} = 0$. This is done by adding an further

Statistica Sinica: Newly accepted Paper

condition (C4) on the positive-definite function φ in (3.1).

(C4) (Condition on φ). $\varphi : [0,1]^2 \to \mathbb{R}$ is a continuous positive-definite symmetric function, such that for any two distinct probability measures μ_1 and μ_2 satisfying $\mu_1((0,1)) = \mu_2((0,1)) = 1$, the set $\mathcal{T}_{\mu_1,\mu_2} := \big\{t \in [0,1] : \mathrm{E}_{Z\sim\mu_1}\{\varphi(t,Z)\} \neq \mathrm{E}_{Z\sim\mu_2}\{\varphi(t,Z)\}\big\}$ is non-empty.

Accordingly, condition (C3) of h is upgraded to the following condition (C5).

(C5) (Equivalence condition on h for achieving (P1)–(P3)). There exists a bivariate function $\varphi : [0,1]^2 \to \mathbb{R}$ satisfying condition (C4), such that $h(y,z) = 2^{-1}\{\varphi(y,y) + \varphi(z,z)\} - \varphi(y,z)$ holds for all $y, z \in [0,1]$.

Theorem 4 verifies that condition (C5) is the necessary and sufficient condition on h such that $\zeta^{(h,F)}$ achieves all properties (P1)–(P3).

Theorem 4 (Necessary and sufficient condition for all properties (P1)–(P3))**.** *Assume that h and F satisfy the regularity conditions (C1) and (C2). Then the correlation measure $\xi^{(h,F)}$ satisfies properties (P1), (P2), and (P3), if and only if h satisfies condition (C5).*

To satisfy condition (C4), φ needs to be a positive-definite function such that the mapping from a probability law μ to the univariate function

Statistica Sinica: Newly accepted Paper

$\mathrm{E}_{Z\sim\mu}\varphi(\cdot, Z)$ is one-to-one. This is equivalent to saying that φ is a characteristic kernel function (Sriperumbudur et al., 2011), as defined in the context of the reproducing kernel Hilbert space (RKHS). Examples of φ satisfying condition (C4) includes:

(a) $\varphi(y, z) = |y|^\gamma + |z|^\gamma - |y - z|^\gamma$, for $\gamma \in (0, 2)$, which gives $h(y, z) = |y - z|^\gamma$;

(b) $\varphi(y, z) = \exp(-\beta\,|y - z|)$, for $\beta > 0$, which yields $h(y, z) = 1 - \exp(-\beta\,|y - z|)$.

(c) $\varphi(y, z) = \exp\{-\theta\,(y - z)^2\}$, for $\theta > 0$, which gives $h(y, z) = 1 - \exp\{-\theta\,(y - z)^2\}$.

Refer to Fukumizu et al. (2008) for a variety of other φ functions that satisfy condition (C4).

To provide a concise overview, Corollary 2 summarizes the results in the previous Theorems 2, 4, and Proposition 1.

Corollary 2 (Summary of results)**.** *Assume that h and F satisfy the regularity conditions (C1) and (C2). We have the following results.*

(i) $\xi^{(h,F)}$ *possesses properties (P2) and (P3)′.*

(ii) $\xi^{(h,F)}$ *possesses properties (P1), (P2) and (P3)′, if and only if h satisfies condition (C3).*

Statistica Sinica: Newly accepted Paper

(iii) $\xi^{(h,F)}$ *possesses properties* (P1), (P2) *and* (P3), *if and only if* h *satisfies condition* (C5).

Remark 3.3. To illustrate each result in Corollary 2, we showcase the properties of $\xi^{(h,F)}$ with $h(y,z) = |y-z|^\gamma$, for $\gamma \in (0,\infty)$. Case (1): $\gamma \in (0,2)$. Then h satisfies condition (C5) with $\varphi(y,z) = |y|^\gamma + |z|^\gamma - |y-z|^\gamma$ satisfying condition (C4). Thus, $\xi^{(h,F)}$ possesses all properties (P1), (P2) and (P3). Case (2): $\gamma = 2$. Then h satisfies condition (C3), with $\varphi(y,z) = yz$. Since this φ does not satisfy condition (C4), $\xi^{(h,F)}$ possesses properties (P1), (P2) and (P3)′, but violates (P3). Case (3): $\gamma \in (2,\infty)$. Then h does not satisfy condition (C3), because $\varphi(y,z) = |y|^\gamma + |z|^\gamma - |y-z|^\gamma$ with $\gamma \in (2,\infty)$ is not positive-definite. Thus, $\xi^{(h,F)}$ possesses properties (P2) and (P3)′, but violates (P1) and (P3).

Remark 3.4. In the special case of $h(y,z) = (y-z)^2$, $\xi^{(h,F)}$ in (2.2) has a simplified expression: $\xi^{(h,F)} = \mathrm{var}[\mathrm{E}\{F(Y) \mid X\}]/\mathrm{var}\{F(Y)\}$. Therefore, $\xi^{(h,F)} = 0$ is equivalent to $\mathrm{E}\{F(Y) \mid X\}$ being invariant in X, which is obviously a weaker condition than X and Y being independent. Thus non-independent X and Y can also result in $\xi^{(h,F)} = 0$, leading to the violation of (P3) for $\xi^{(h,F)}$ with $h(y,z) = (y-z)^2$.

Statistica Sinica: Newly accepted Paper

4. Rank-based correlation coefficients

As previously noted, the rank-based Chatterjee's correlation coefficient (1.1) corresponds to our coefficient $\xi_n^{(h,F)}(X,Y)$ when taking $h(y,z) = |y-z|$ and $F = F_{Y,n}$ as the empirical c.d.f. of Y. This motivates us to investigate the rank-based correlation coefficient $\widetilde{\xi}_n^{(h)}(X,Y)$ in Definition 2, which is a variant of the previous $\xi_n^{(h,F)}(X,Y)$ when letting $F = F_{Y,n}$.

Definition 2 (Rank-based correlation coefficient)**.** For $y \in \mathbb{R}$, let $R(y) = nF_{Y,n}(y) = \sum_{j=1}^{n} \mathrm{I}(Y_j \le y)$ be the rank of y among all Y_j's. For a non-negative continuous bivariate function $h : [0,1]^2 \to [0,\infty)$ satisfying $h(x,x) = 0$ for all $x \in [0,1]$, define the variation statistic

$$\widetilde{\zeta}_n^{(h)}(X,Y) := \frac{1}{n}\sum_{i=1}^{n-1} h\Big\{\frac{R(Y_{[i]})}{n}, \frac{R(Y_{[i+1]})}{n}\Big\}, \tag{4.1}$$

and the normalization statistic

$$\widetilde{\chi}_n^{(h)}(Y) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h\Big\{\frac{R(Y_i)}{n}, \frac{R(Y_j)}{n}\Big\}. \tag{4.2}$$

The proposed rank-based correlation coefficient is defined as

$$\widetilde{\xi}_n^{(h)}(X,Y) \;:=\; 1 - \widetilde{\zeta}_n^{(h)}(X,Y)/\widetilde{\chi}_n^{(h)}(Y)$$

Statistica Sinica: Newly accepted Paper

$$= 1 - \frac{n\sum_{i=1}^{n-1} h\{n^{-1}R(Y_{[i]}), n^{-1}R(Y_{[i+1]})\}}{\sum_{i=1}^{n}\sum_{j=1}^{n} h\{n^{-1}R(Y_i), n^{-1}R(Y_j)\}}.$$

In case that $\widetilde{\chi}_n^{(h)}(Y) = 0$, we set $\widetilde{\xi}_n^{(h)}(X,Y) = 1$.

It is noteworthy that all the theorems in previous sections are mainly derived for $\xi_n^{(h,F)}(X,Y)$ with deterministic h and F, while the properties of the rank-based coefficient $\widetilde{\xi}_n^{(h)}(X,Y) = \xi_n^{(h,F_{Y,n})}(X,Y)$ with non-deterministic $F_{Y,n}$ remains unexplored. In our context, it is important to note that the empirical c.d.f. $F_{Y,n}$ differs fundamentally from a typical c.d.f. F in Definition 1. First, from an asymptotic standpoint, $F_{Y,n}$ is a random function depending on the sample Y_i's and sample size n, whereas F is deterministic and independent from Y. This disparity significantly impacts the derivation of asymptotic theories. Second, $F_{Y,n}$ in its nature is a piecewise-constant function, which obviously violates the continuity assumption for F in condition (C2). A related result is Corollary 1, which establishes the convergence of the correlation coefficient for a general random $\widehat{F}_n$. However, its applicability requires h and $\widehat{F}_n$ to satisfy certain additional conditions. For these reasons, the asymptotic properties of $\widetilde{\xi}_n^{(h)}(X,Y)$ can not be directly inferred from the previous theorems, but need to be investigated separately and specifically.

Theorem 5 shows that $\widetilde{\xi}_n^{(h)}(X,Y)$ also converges almost surely to a

Statistica Sinica: Newly accepted Paper

deterministic limit, provided that function h is Lipschitz-continuous.

(C6) (Lipschitz continuity condition on h). There exists a constant $K_h \in (0, \infty)$, such that $|h(y_1, z_1) - h(y_2, z_2)| \leq K_h \left(|y_1 - y_2| + |z_1 - z_2|\right)$, for any $(y_1, z_1), (y_2, z_2) \in [0, 1]^2$.

Theorem 5 (Convergence of rank-based correlation coefficient)**.** *Assume that h satisfies conditions (C1) and (C6). As $n \to \infty$, we have the following three convergence results:* (i) *the variation statistic $\widetilde{\zeta}_n^{(h,F)}(X, Y)$ converges a.s. to the limit*

$$\widetilde{\zeta}^{(h)} := \int\int\int h\big\{F_Y(y), F_Y(z)\big\}\, \mathrm{d}\mu_{Y|x}(y)\, \mathrm{d}\mu_{Y|x}(z)\, \mathrm{d}\mu_X(x);$$

(ii) *the normalization statistic $\widetilde{\chi}_n^{(h)}(Y)$ converges a.s. to the strictly-positive limit*

$$\widetilde{\chi}^{(h)} := \int\int h\{F_Y(y), F_Y(z)\}\, \mathrm{d}\mu_Y(y)\, \mathrm{d}\mu_Y(z); \tag{4.3}$$

and (iii) *the correlation coefficient $\widetilde{\xi}_n^{(h)}(X, Y)$ converges a.s. to the limit*

$$\widetilde{\xi}^{(h)} := 1 - \widetilde{\zeta}^{(h)}/\widetilde{\chi}^{(h)}. \tag{4.4}$$

Moreover, when X and Y are independent, the central limit theorem

Statistica Sinica: Newly accepted Paper

also holds for $\widetilde{\xi}_n^{(h)}(X,Y)$.

Theorem 6 (CLT of $\widetilde{\xi}_n^{(h)}(X,Y)$ under independence)**.** *Assume that h satisfies conditions (C1) and (C6). Assume that X and Y are independent. Then as $n \to \infty$, we have*

$$\sqrt{n}\,\widetilde{\xi}_n^{(h)}(X,Y) \xrightarrow{\mathcal{D}} N(0,\widetilde{\sigma}^2),$$

where $\widetilde{\sigma}^2$ is obtained by plugging $F = F_Y$ into the expression of σ^2 in Theorem 3.

Remark 4.1. It is a well-known fact that $F_Y(Y)$ follows a Unif$[0,1]$ distribution if Y is continuous. In such case, $\widetilde{\sigma}^2$ reduces to

$$\big[\mathrm{E}\big\{h(U_1,U_2)^2\big\} - 2\mathrm{E}\big\{h(U_1,U_2)h(U_1,U_3)\big\} + \mathrm{E}\big\{h(U_1,U_2)\big\}^2\big]/\mathrm{E}\big\{h(U_1,U_2)\big\}^2,$$

with $U_1, U_2, U_3 \sim_{\text{i.i.d.}} \text{Unif}[0,1]$. Owing to this expression, $\widetilde{\sigma}^2$ can be directly computed via numerical integrals, without the need of being estimated from the observed data. For example, when taking $h(y,z) = |y-z|^\gamma$ with $\gamma > 0$, we have

$$\widetilde{\sigma}^2 = 1 + (\gamma+2)^2\left\{\frac{\gamma+1}{4(2\gamma+1)} - \frac{1}{2\gamma+3} - \frac{\Gamma(\gamma+2)^2}{\Gamma(2\gamma+4)}\right\},$$

Statistica Sinica: Newly accepted Paper

where $\Gamma(\cdot)$ represents the gamma function. Letting $\gamma = 1$ yields $\widetilde{\sigma}^2 = 2/5$, which agrees with the variance τ^2 in Chatterjee (2021) [Theorem 2.2].

4.1 Simplified rank-based coefficient under continuous Y

Note that $F_Y(Y) \sim \text{Unif}[0,1]$ for any continuous random variable Y. Therefore, if Y is continuous, the denominator term $\widetilde{\chi}^{(h)}$ in (4.3) could be simplified as

$$\widetilde{\chi}^{(h)} = \int_0^1 \int_0^1 h(u,v)\, \mathrm{d}u\, \mathrm{d}v =: C_h, \tag{4.5}$$

which is a positive constant that does not rely on the distribution law of Y. In this scenario, the normalization statistic $\widetilde{\chi}_n^{(h)}(Y)$ in (4.2), acting as consistent estimator for $\widetilde{\chi}^{(h)}$, could be simply replaced by the exact value $C_h = \widetilde{\chi}^{(h)}$. This gives rises to a simplified version of the rank-based correlation coefficient, as presented in Definition 3.

Definition 3 (Simplified rank-based correlation coefficient under continuous Y)**.** Assume that Y is continuous. For a non-negative non-constant continuous bivariate function $h : [0,1]^2 \to [0,\infty)$ satisfying $h(x,x) = 0$ for

Statistica Sinica: Newly accepted Paper

all $x \in [0,1]$, define the simplified rank-based correlation coefficient

$$\begin{aligned} \breve{\xi}_n^{(h)}(X,Y) &:= 1 - \widetilde{\zeta}_n^{(h)}(X,Y)/C_h \\ &= 1 - \frac{n^{-1}\sum_{i=1}^{n-1} h\big\{n^{-1}R(Y_{[i]}), n^{-1}R(Y_{[i+1]})\big\}}{\int_0^1\int_0^1 h(u,v)\,\mathrm{d}u\,\mathrm{d}v}, \end{aligned}$$

where $\widetilde{\zeta}_n^{(h)}(X,Y)$ and C_h are defined in (4.1) and (4.5) respectively.

A major appealing property for $\breve{\xi}_n^{(h)}(X,Y)$ is its reduced computational cost compared to that of the original $\widetilde{\xi}_n^{(h)}(X,Y)$. This is because the normalization constant C_h in $\breve{\xi}_n^{(h)}(X,Y)$ does not rely on Y and could be directly computed from the integral (4.5). In fact, for many common h functions, C_h has simple close-form expressions. Take our previous examples of h in Section 3.2, if $h(y,z) = |y-z|^\gamma$ for $\gamma > 0$, then $C_h = 2(\gamma+1)^{-1}(\gamma+2)^{-1}$; and if $h(y,z) = 1-\exp(-\beta|y-z|)$ for $\beta > 0$, then $C_h = 1 - 2\beta^{-1} + 2\beta^{-2} - 2\beta^{-2}\exp(-\beta)$. Corollary 3 ensures that $\breve{\xi}_n^{(h)}(X,Y)$ preserves all the asymptotic properties of $\widetilde{\xi}_n^{(h)}(X,Y)$ in Theorems 5 and 6.

Corollary 3 (Consistency and CLT for simplified correlation coefficient)**.** *Assume Y is continuous. Assume h satisfies conditions (C1) and (C6). Then as $n \to \infty$, $\breve{\xi}_n^{(h)}(X,Y)$ converges a.s. to the limit $\widetilde{\xi}^{(h)}$ in (4.4). Moreover, if X and Y are independent, then the central limit theorem holds: $\sqrt{n}\breve{\xi}_n^{(h)}(X,Y) \xrightarrow{\mathcal{D}} N(0,\widetilde{\sigma}^2)$ as $n \to \infty$, where $\widetilde{\sigma}^2$ is defined in Theorem 6.*

Statistica Sinica: Newly accepted Paper

To conclude, $\breve{\xi}_n^{(h)}(X,Y)$ achieves the same nice asymptotic properties as $\widetilde{\xi}_n^{(h)}(X,Y)$, but with much reduced computational expense. Therefore, $\breve{\xi}_n^{(h)}(X,Y)$ is a more efficient alternative to $\widetilde{\xi}_n^{(h)}(X,Y)$, if knowing that Y_i's take values in continuous support.

4.2 Properties of correlation measure $\widetilde{\xi}^{(h)}$

This subsection investigates the properties of correlation measure $\widetilde{\xi}^{(h)}$, which is the limit of the rank-based correlation coefficient $\widetilde{\xi}_n^{(h)}(X,Y)$ or its simplified version $\breve{\xi}_n^{(h)}(X,Y)$. A comparison between the correlation measures $\widetilde{\xi}^{(h)}$ in (4.4) and $\xi^{(h,F)}$ in (2.2) gives that $\widetilde{\xi}^{(h)} = \xi^{(h,F_Y)}$, that is, $\widetilde{\xi}^{(h)}$ coincides with $\xi^{(h,F)}$ when substituting $F = F_Y$. One may speculate that $\widetilde{\xi}^{(h)}$ also preserves similar properties of $\xi^{(h,F)}$ as presented in Theorems 2, 4, and Proposition 1. Again, this needs to be reexamined, as we should not simply regard $\widetilde{\xi}^{(h)} = \xi^{(h,F_Y)}$ as a special case of $\xi^{(h,F)}$. For all previous theorems related to $\xi^{(h,F)}$, F is assumed to be a fixed function that does not depend on the (X,Y) distribution and satisfies the regularity condition (C2). In contrast, F_Y not only depends on the (X,Y) distribution, but also possibly violates the required regularity condition (C2). We stress that our investigation into properties (P1)–(P3) encompasses all possible probability distributions of (X,Y), only except for constant X or Y.

Statistica Sinica: Newly accepted Paper

To proceed, we first present conditions (C4)′ and (C5)′, which are modified version of conditions (C4) and (C5) in Section 3.2.

(C4)′ (Modified condition on φ). $\varphi : [0,1]^2 \to \mathbb{R}$ is a continuous positive-definite symmetric function, such that for any two distinct probability measures μ_1 and μ_2 satisfying $\mu_1((0,1]) = \mu_2((0,1]) = 1$, the set $\mathcal{T}_{\mu_1,\mu_2} := \left\{t \in [0,1] : \mathrm{E}_{Z\sim\mu_1}\{\varphi(t,Z)\} \neq \mathrm{E}_{Z\sim\mu_2}\{\varphi(t,Z)\}\right\}$ is non-empty.

(C5)′ (Modified equivalence condition on h for achieving (P1)–(P3)). There exists some bivariate function $\varphi : [0,1]^2 \to \mathbb{R}$ satisfying condition (C4)′, such that $h(y,z) = 2^{-1}\{\varphi(y,y) + \varphi(z,z)\} - \varphi(y,z)$ holds for all $y, z \in [0,1]$.

Note that the difference between conditions (C4)′ and (C4) is minor: the open interval $(0,1)$ in condition (C4) is modified to be the wider half-open interval $(0,1]$ in condition (C4) (i.e., "$\mu_1((0,1)) = \mu_2((0,1)) = 1$" in (C4) is revised to be "$\mu_1((0,1]) = \mu_2((0,1]) = 1$" in (C4)′). Therefore, the revised conditions (C4)′ and (C5)′ are slightly stronger than the previous conditions (C4) and (C5). Theorem 7 verifies that $\widetilde{\xi}^{(h)}$ possesses similar properties to $\xi^{(h,F)}$.

Statistica Sinica: Newly accepted Paper

Theorem 7 (Properties of correlation measure $\widetilde{\xi}^{(h)}$)**.** *Assume that h satisfies the regularity condition (C1). Then we have the following results.*

(i) $\widetilde{\xi}^{(h)}$ *possesses properties (P2) and (P3)′.*

(ii) $\widetilde{\xi}^{(h)}$ *possesses properties (P1), (P2) and (P3)′, if and only if h satisfies condition (C3).*

(iii) $\widetilde{\xi}^{(h)}$ *possesses properties (P1), (P2) and (P3), if h satisfies condition (C5)′.*

In comparison with the previous Corollary 2 for $\xi^{(h,F)}$, most results in Theorem 7 for $\widetilde{\xi}^{(h)}$ remain unchanged. The only exception is that we solely confirmed condition (C5)′ as the sufficiency condition for attaining all properties (P1)–(P3), leaving uncertain whether it is also the necessary condition. This poses an open question for future resolution.

4.3 Supplementary literature review

We first review several information-theoretic measures related to the present work. It is worth noting that mutual information (Cover and Thomas, 2006), due to the lack of normalization, satisfies only property (P3) but not (P1) or (P2). One of its normalized variants, the uncertainty coefficient (Theil's U) (Theil, 1970), can satisfy all three properties, but this requires

Statistica Sinica: Newly accepted Paper

X and Y to be discrete variables. Similarly, similarity metrics such as the normalized information distance (NID) (Li et al., 2004) and normalized mutual information (NMI) (Kvålseth, 2017) satisfy properties (P1), (P3), and partially (P2), and are also restricted to discrete variables. When applied to continuous or mixed-type random variables, these information-theoretic measures generally require complex transformations and estimation procedures (Darbellay and Vajda, 1999; Kraskov et al., 2004; Gao et al., 2018), often resulting in modifications to their theoretical properties, which represents a common limitation.

There are several recent works that extended the Chatterjee's correlation coefficient ξ_n in different directions. Azadkia and Chatterjee (2021) proposed a correlation coefficient based on ξ_n for measuring the conditional dependence of Y on Z given X, and accordingly designed a new variable selection algorithm. Lin and Han (2022a) proved the CLT for Chatterjee's correlation coefficient under arbitrary distributions of (X, Y). Auddy et al. (2024), Bickel (2022) and Shi et al. (2021) conducted power analyses for ξ_n under a family of alternatives. Lin and Han (2022b) extended ξ_n to enhance power for independence tests, by incorporating multiple nearby ranks. Lin and Han (2024) further proved the bootstrap inconsistency for ξ_n. Generalizations of ξ_n to multi-dimensional spaces were proposed in Deb et al.

Statistica Sinica: Newly accepted Paper

(2020), Huang et al. (2022), Gamboa et al. (2022), and Fuchs (2024). Scientific applications of ξ_n across various fields could be found in Sadeghi (2022), Zarei et al. (2023), and Nurdiati et al. (2022). Refer to Chatterjee (2024) for a comprehensive review of recent follow-up works. While these works focus on various topics such as independence test, power analysis, variable selection, and high-dimensional extensions, our research takes a distinct approach. We concentrate on tackling the fundamental task of quantifying functional dependence between two random scalars — an objective that aligns with the original intention of the Chatterjee's work.

Among these related works, Deb et al. (2020) proposed a kernel measure of association (KMAc) for multivariate X and Y, which was subsequently extended to conditional correlation by Huang et al. (2022). These two works are closely related to our framework. As shown in Section 3, when the function h in our correlation measure $\xi^{(h,F)}$ satisfies condition (C3), namely, when h admits a decomposition in terms of kernels φ, $\xi^{(h,F)}$ can essentially be viewed as a scalar special case of KMAc. Our framework, however, starts from a general distance-generating function h rather than a kernel-based construction, and thus also covers a broad class of cases where h does not satisfy condition (C3), leading to the new results in Theorems 1–3. These results are non-trivial and can not be directly obtained through kernel-based

Statistica Sinica: Newly accepted Paper

techniques. Moreover, while Deb et al. (2020) established the sufficiency of the kernel-based construction for properties (P1)–(P3), we characterize the necessity of condition (C3), showing that, within the general class of h, the three properties can be simultaneously satisfied only when h admits the corresponding kernel decomposition. Finally, the introduction of the c.d.f. F provides a unified formulation encompassing both the non-rank-based measure $\xi^{(h,F)}$ and the rank-based measures $\widetilde{\xi}^{(h)} = \xi^{(h,F_{Y,n})}$ and $\breve{\xi}^{(h)}$. Although these measures share a common formulation, their theoretical properties require separate treatment, leading to our new theoretical results established in Theorems 5-7.

5. Simulation studies

In this section, we evaluate the performance of our proposed correlation coefficients via simulations on synthetic data.

5.1 Set-up

For generating the synthetic data $\{(X_i, Y_i)\}_{i=1,\ldots,n}$, we adopt the following three models as used in section 4 of Chatterjee (2021):

Model 1 (Linear): $Y = X + \theta E$.

Model 2 (Quadratic): $Y = X^2 + \theta E$.

Statistica Sinica: Newly accepted Paper

Model 3 (Sinusoidal): $Y = \sin(2\pi X) + \theta E$.

In each of the three models mentioned above, we set $X \sim \text{Unif}[-1, 1]$. Additionally, $E \sim N(0, 1)$ is the Gaussian noise independent of X, and $\theta \in [0, \infty]$ represents the noise magnitude. With a slight abuse of notation, we also allow taking $\theta = \infty$, which corresponds to the scenario where $Y = E$ is pure noise.

We assess the performance of the correlation coefficient $\xi_n^{(h,F)}$ in Definition 1 with $F = \Phi$ being the c.d.f. of $N(0, 1)$ distribution, as well as the simplified rank-based correlation coefficient $\breve{\xi}_n^{(h)}$ in Definition 3. For both of $\xi_n^{(h,F)}$ and $\breve{\xi}_n^{(h)}$, the bivariate function h is specified from the following five cases: (1) $h_1(y, z) = |y - z|$; (2) $h_2(y, z) = |y - z|^2$; (3) $h_3(y, z) = |y - z|^3$; (4) $h_4(y, z) = 1 - e^{|y-z|}$; and (5) $h_5(y, z) = 1 - e^{(y-z)^2}$. According to Corollary 1 and Theorem 7, the correlation measures $\xi^{(h,F)}$ and $\widetilde{\xi}^{(h)}$ can possess different properties of (P1)–(P3) contingent upon the specific h functions. Specifically, h_1, h_4, and h_5 lead to properties (P1), (P2), and (P3); h_2 and results in properties (P1), (P2), and (P3)′; and h_3 yields properties (P2), and (P3)′. These h functions all give rise to the perfect dependence property (P2), indicating the validity of the corresponding correlation coefficients for detecting arbitrarily-shaped functional dependence.

Statistica Sinica: Newly accepted Paper

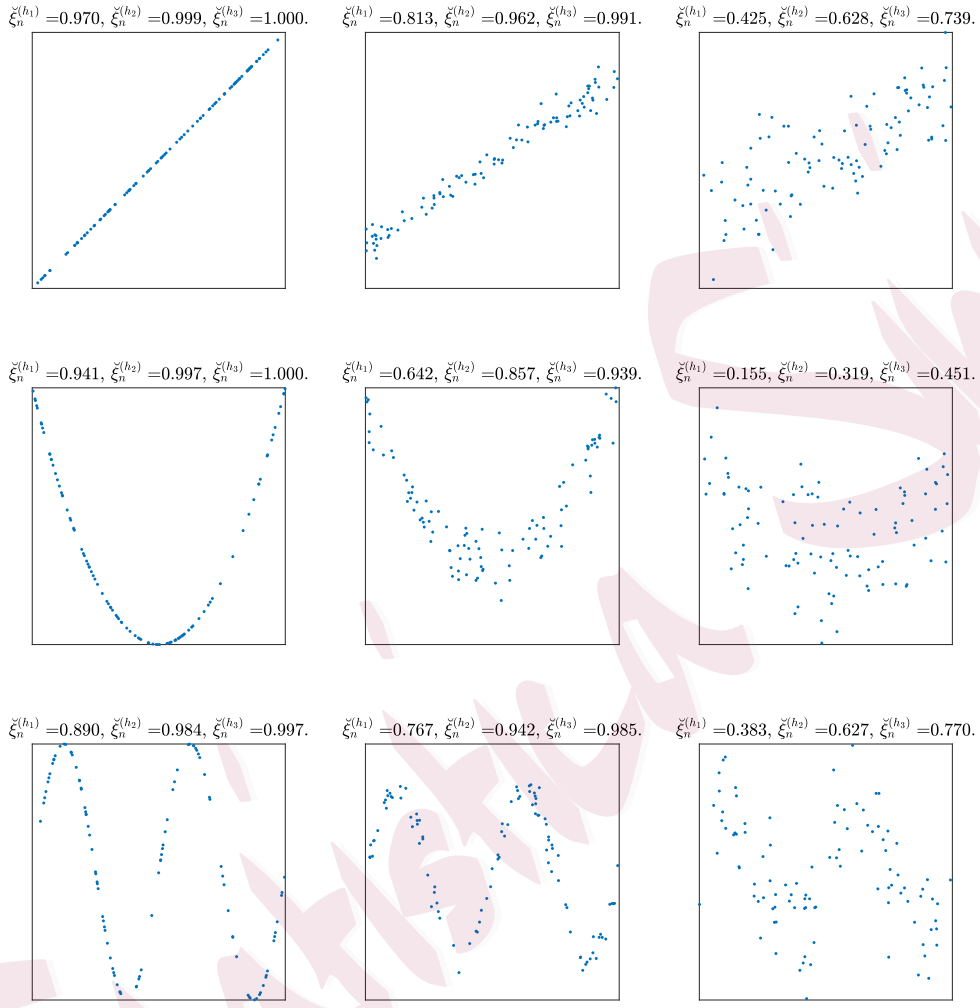

Figure 1: Illustrative example for the simplified rank-based correlation coefficients $\breve{\xi}_n^{(h)}$, with $h_1 = |y - z|$, $h_2 = |y - z|^2$ and $h_3 = |y - z|^3$. In the 3-by-3 panel grids, rows from top to bottom correspond to Models 1, 2, and 3; and columns from left to right correspond to noise level $\theta = 0, 0.1$ and 0.5. The sample size $n = 100$. Note that $\breve{\xi}_n^{(h_1)}$ coincides with the Chatterjee's correlation coefficient.

Statistica Sinica: Newly accepted Paper

5.2 Illustrative example

Figure 1 provides an illustrative example on the performance of simplified rank-based correlation coefficient $\breve{\xi}_n^{(h)}$, with $h = h_1$, h_2 or h_3. Note that $\breve{\xi}_n^{(h)}$ with $h_1 = |y - z|$ coincides with the Chatterjee's correlation coefficient in (1.1), which serves as a benchmark for method comparisons. Across all models, the three correlation coefficients exhibit higher values when the noise level is lower. In each panel of Figure 1, the three correlation coefficients consistently follow this order: $\breve{\xi}_n^{(h_1)} < \breve{\xi}_n^{(h_2)} < \breve{\xi}_n^{(h_3)}$. This suggests that $\breve{\xi}_n^{(h_2)}$ and $\breve{\xi}_n^{(h_3)}$ are more sensitive than $\breve{\xi}_n^{(h_1)}$ in identifying functional dependence, especially under the case of winding-shaped Models 2 and 3. In particular, for the noise-free case $\theta = 0$, all three correlation coefficients are expected to converge to 1 according to the perfect dependence property (P2). It is observed that $\breve{\xi}_n^{(h_2)}$ and $\breve{\xi}_n^{(h_3)}$ are much closer to 1 (with the largest difference < 0.04) than $\breve{\xi}_n^{(h_1)}$ (with the largest difference > 0.1). This indicates that $\breve{\xi}_n^{(h_2)}$ and $\breve{\xi}_n^{(h_3)}$ have higher convergence speed than $\breve{\xi}_n^{(h_1)}$, or in other words, $\breve{\xi}_n^{(h_2)}$ and $\breve{\xi}_n^{(h_3)}$ can attain enhanced detection results with a reduced sample size requirement. This simulation example reveals that among our proposed family of correlation coefficients, some $\breve{\xi}_n^{(h)}$ with appropriately-chosen h functions may have better practical abilities for detecting functional relationship than the benchmark Chatterjee's correlation

Statistica Sinica: Newly accepted Paper

coefficient.

5.3 Comparison with other correlation coefficients

To provide a more comprehensive evaluation, we compare our correlation coefficient with some other methods, which includes the Pearson's correlation coefficient, the Spearman's ρ, the distance correlation (abbreviated as "dCorr") in Székely et al. (2007), the power-enhanced Chatterjee's correlation coefficient in Lin and Han (2022b) (abbreviated as "LH", with the number of nearest neighbor $M = 10$), the Maximal Information Coefficient (MIC) in Reshef et al. (2011), the Mutual Information (MI), and the uncertainty coefficient (Theil's U) in Theil (1970). The mutual information (MI) is estimated from the observed samples using the KSG method (Kraskov et al., 2004) with the number of nearest neighbors set to $k = 5$. For estimating Theil's U, we first discretize the data $(X_i, Y_i)_{i=1}^n$ by partitioning both the x- and y-axes into $m = \lceil \sqrt{n} \rceil$ intervals. Based on this discretization, we then compute the empirical mutual information $I(X, Y)$ and the empirical entropy $H(Y)$, and obtain Theil's U $= I(X, Y)/H(Y)$. This procedure ensures that the resulting Theil's U values lie within the $[0, 1]$ interval, thereby achieving a proper normalization that facilitates a fair comparison with our proposed method.

Statistica Sinica: Newly accepted Paper

Under the linear Model 1, all methods exhibit similar performances; therefore, the results are omitted. Tables 1 and 2 report the values of correlation coefficients of each method under Models 2 and 3 respectively. As the sample size n grows from 100 to 2000, the mean values of each correlation coefficients slightly change, whereas the standard deviations become closer to zero. This implies a trend of asymptotic convergence for $\xi_n^{(h,F)}$ and $\breve{\xi}_n^{(h)}$, which agrees with our Theorems 1 and 5. Under both the two models, our proposed correlation coefficients $\xi_n^{(h,F)}$ and $\breve{\xi}_n^{(h)}$ distinguish themselves from other methods by clearly reflecting the amount of functional dependence between X and Y. It is seen that $\xi_n^{(h,F)}$ and $\breve{\xi}_n^{(h)}$ are approximately equal to 1 under perfect dependence (i.e., $\theta = 0$), then decrease as the noise level θ grows, and finally approaches to zero when $Y = E$ (i.e., $\theta = \infty$) consists of pure noise. The MIC shows a similar trend, but it does not approach zero closely when the noise level goes to infinity, as it has the second largest value among all methods under the case of $\theta = \infty$. Though the MIC is theoretically guaranteed to converge to zero (i.e., satisfying property (P3)′) for independent X and Y, our simulation results indicate that its convergence speed is significantly slower than other methods. The LH correlation coefficient is comparable to the Chatterjee's correlation coefficient $\breve{\xi}_n^{(h_1)}$ when the sample size is large ($n \in \{500, 2000\}$), but exhibits

Statistica Sinica: Newly accepted Paper

inferior performance when the sample size is small ($n = 100$). The distance correlation "dCorr" shows a trend of decreasing to zero as θ grows, while it is far away from reaching 1 for the noise-free case ($\theta = 0$). The mutual information (MI) also decreases rapidly toward zero as the noise level θ grows; however, at $\theta = 0$, its behavior is highly unstable. For instance, under Model 3, when the sample size n increases from 100 to 2000, the MI value surges from 0.724 to 3.34. This instability highlights the weakness of MI in detecting perfect functional dependence. Theil's U likewise shows a decreasing trend toward zero as θ grows. Nevertheless, it stays noticeably below 1 at $\theta = 0$ and remains well above 0 at $\theta = \infty$, which deviates considerably from the theoretically desired properties (namely, satisfying properties (P1)–(P3)). This discrepancy is likely due to the discretization step involved in estimating Theil's U, which introduces larger estimation errors and slower convergence, thereby reflecting its limitations in practical applications. The Pearson correlation coefficient and Spearman's ρ fail to identify significant functional relationships in both models, as they are either negative or close to zero even under strong functional dependence ($\theta \in \{0, 0.1\}$). For the comparison between the two proposed coefficient $\xi_n^{(h,F)}$ and $\breve{\xi}_n^{(h)}$, they exhibit similar values when using identical h functions. Among the five selected h functions, h_2, h_3 and h_5 yield higher values of

Statistica Sinica: Newly accepted Paper

correlation coefficients than h_1 and h_4 do, especially when the functional dependence is strong ($\theta \in \{0, 0.1, 0.5\}$). This also agrees with the previous results reflected in Figure 1.

Table 1: Simulation results under Model 2: $Y = X^2 + \theta E$. Experiments are replicated for 100 times, with mean values of correlation coefficients shown in the table, and standard deviations (multiplied by 100) presented in parentheses. For $\xi_n^{(h,F)}$, we use $F = \Phi$, the c.d.f. of standard normal distribution.

	Method	$\theta = 0$, $(Y = f(X))$			$\theta = 0.1$			$\theta = 0.5$			$\theta = \infty$, $(X \perp Y)$		
		n=100	n=500	n=2000	n=100	n=500	n=2000	n=100	n=500	n=2000	n=100	n=500	n=2000
$\xi_n^{(h,F)}$	$h_1 = \lvert y-z \rvert$	0.943	0.989	0.997	0.674	0.680	0.680	0.126	0.145	0.144	0.004	-0.003	0.000
		(0.25)	(0.03)	(0.00)	(3.20)	(1.66)	(0.87)	(6.26)	(2.94)	(1.63)	(6.16)	(2.62)	(1.39)
	$h_2 = \lvert y-z \rvert^2$	0.995	1.000	1.000	0.890	0.893	0.894	0.221	0.251	0.250	-0.000	-0.003	-0.000
		(0.08)	(0.00)	(0.00)	(2.00)	(1.03)	(0.54)	(10.01)	(4.49)	(2.40)	(9.57)	(4.07)	(2.16)
	$h_3 = \lvert y-z \rvert^3$	0.999	1.000	1.000	0.961	0.963	0.962	0.296	0.332	0.332	-0.008	-0.002	-0.001
		(0.03)	(0.00)	(0.00)	(1.04)	(0.54)	(0.29)	(13.22)	(5.81)	(2.92)	(12.90)	(5.53)	(2.85)
	$h_4 = 1 - e^{\lvert y-z \rvert}$	0.939	0.988	0.997	0.653	0.659	0.659	0.111	0.128	0.128	0.005	-0.003	0.000
		(0.25)	(0.03)	(0.00)	(3.27)	(1.70)	(0.89)	(5.77)	(2.73)	(1.52)	(5.53)	(2.36)	(1.24)
	$h_5 = 1 - e^{(y-z)^2}$	0.995	1.000	1.000	0.887	0.891	0.891	0.209	0.237	0.236	0.003	-0.003	0.000
		(0.09)	(0.00)	(0.00)	(2.04)	(1.05)	(0.55)	(9.57)	(4.31)	(2.34)	(8.68)	(3.70)	(1.98)
$\xi_n^{(h)}$	$h_1 = \lvert y-z \rvert$ (Chatterjee)	0.941	0.988	0.997	0.645	0.660	0.664	0.134	0.155	0.155	0.004	-0.003	0.000
		(0.04)	(0.00)	(0.00)	(3.50)	(1.94)	(0.91)	(6.66)	(3.07)	(1.73)	(6.10)	(2.63)	(1.38)
	$h_2 = \lvert y-z \rvert^2$	0.997	1.000	1.000	0.859	0.865	0.869	0.225	0.257	0.257	-0.000	-0.003	0.000
		(0.03)	(0.00)	(0.00)	(2.90)	(1.59)	(0.71)	(10.27)	(4.47)	(2.54)	(9.52)	(4.08)	(2.16)
	$h_3 = \lvert y-z \rvert^3$	1.000	1.000	1.000	0.938	0.941	0.943	0.292	0.331	0.331	-0.008	-0.002	-0.001
		(0.01)	(0.00)	(0.00)	(2.04)	(1.08)	(0.47)	(12.85)	(5.48)	(3.05)	(12.84)	(5.54)	(2.84)
	$h_4 = 1 - e^{\lvert y-z \rvert}$	0.927	0.985	0.996	0.593	0.609	0.614	0.113	0.132	0.132	0.005	-0.003	0.000
		(0.05)	(0.00)	(0.00)	(3.65)	(2.02)	(0.96)	(5.93)	(2.79)	(1.56)	(5.48)	(2.36)	(1.23)
	$h_5 = 1 - e^{(y-z)^2}$	0.996	1.000	1.000	0.837	0.844	0.848	0.203	0.233	0.234	0.002	-0.003	0.000
		(0.03)	(0.00)	(0.00)	(3.21)	(1.77)	(0.80)	(9.60)	(4.22)	(2.42)	(8.65)	(3.71)	(1.98)
	Pearson	-0.005	-0.007	0.003	0.021	0.007	0.005	-0.004	-0.006	-0.004	0.003	0.007	-0.000
		(12.98)	(5.04)	(2.42)	(12.10)	(5.83)	(2.72)	(11.08)	(4.51)	(2.39)	(8.65)	(4.75)	(2.33)
	Spearman	-0.003	-0.007	0.000	0.010	0.002	0.005	-0.004	-0.007	-0.003	0.003	0.009	-0.001
		(14.46)	(5.64)	(2.67)	(12.91)	(5.96)	(3.27)	(12.02)	(4.81)	(2.38)	(8.63)	(4.65)	(2.41)
	dCorr	0.514	0.496	0.492	0.485	0.468	0.462	0.278	0.245	0.239	0.160	0.074	0.038
		(2.82)	(1.04)	(0.46)	(2.65)	(1.05)	(0.53)	(4.37)	(1.75)	(0.91)	(3.15)	(1.68)	(0.90)
	MI	1.804	3.303	4.647	0.941	1.054	1.053	0.129	0.156	0.154	0.023	0.011	0.006
		(6.53)	(3.66)	(1.69)	(8.27)	(4.31)	(2.24)	(6.80)	(3.40)	(1.61)	(3.26)	(1.68)	(0.87)
	Theil's U	0.623	0.721	0.768	0.454	0.396	0.342	0.287	0.245	0.209	0.256	0.222	0.194
		(2.69)	(0.78)	(0.35)	(2.50)	(0.98)	(0.43)	(2.95)	(1.08)	(0.48)	(2.23)	(0.78)	(0.33)
	LH	0.611	0.918	0.979	0.470	0.638	0.660	0.026	0.134	0.150	-0.161	-0.032	-0.008
		(0.00)	(0.00)	(0.00)	(2.45)	(1.59)	(0.75)	(5.18)	(2.05)	(1.28)	(2.62)	(0.95)	(0.46)
	MIC	1.000	1.000	1.000	0.876	0.847	0.813	0.342	0.285	0.240	0.238	0.163	0.107
		(0.00)	(0.00)	(0.00)	(6.26)	(2.96)	(1.57)	(6.27)	(2.63)	(1.59)	(3.62)	(1.40)	(0.51)

For a more intuitive presentation of results, Figure 2 visualizes the trend of correlation coefficient with growing noise magnitude θ for the seven

Statistica Sinica: Newly accepted Paper

Table 2: Simulation results under Model 3: $Y = \sin(2\pi X) + \theta E$. Experiments are replicated for 100 times, with mean values of correlation coefficients shown in the table, and standard deviations (multiplied by 100) presented in parentheses. For $\xi_n^{(h,F)}$, we use $F = \Phi$, the c.d.f. of standard normal distribution.

	Method	$\theta = 0,\ (Y = f(X))$			$\theta = 0.1$			$\theta = 0.5$			$\theta = \infty,\ (X \perp Y)$		
		n=100	n=500	n=2000	n=100	n=500	n=2000	n=100	n=500	n=2000	n=100	n=500	n=2000
$\xi_n^{(h,F)}$	$h_1 = \lvert y-z\rvert$	0.907	0.981	0.995	0.840	0.874	0.877	0.462	0.468	0.470	0.004	-0.003	0.000
		(0.37)	(0.03)	(0.00)	(1.13)	(0.56)	(0.30)	(5.37)	(2.50)	(1.35)	(6.16)	(2.62)	(1.39)
	$h_2 = \lvert y-z\rvert^2$	0.985	0.999	1.000	0.970	0.983	0.984	0.686	0.693	0.696	-0.000	-0.003	-0.000
		(0.18)	(0.00)	(0.00)	(0.44)	(0.15)	(0.08)	(6.06)	(2.76)	(1.43)	(9.57)	(4.07)	(2.16)
	$h_3 = \lvert y-z\rvert^3$	0.997	1.000	1.000	0.993	0.997	0.998	0.803	0.809	0.812	-0.008	-0.002	-0.001
		(0.11)	(0.00)	(0.00)	(0.20)	(0.03)	(0.02)	(5.65)	(2.58)	(1.28)	(12.90)	(5.53)	(2.85)
	$h_4 = 1 - e^{\lvert y-z\rvert}$	0.890	0.977	0.994	0.811	0.850	0.852	0.411	0.417	0.419	0.005	-0.003	0.000
		(0.39)	(0.03)	(0.00)	(1.23)	(0.64)	(0.35)	(5.15)	(2.42)	(1.31)	(5.53)	(2.36)	(1.24)
	$h_5 = 1 - e^{(y-z)^2}$	0.983	0.999	1.000	0.966	0.981	0.982	0.655	0.662	0.665	0.003	-0.003	0.000
		(0.20)	(0.00)	(0.00)	(0.49)	(0.16)	(0.09)	(6.22)	(2.82)	(1.48)	(8.68)	(3.70)	(1.98)
$\check{\xi}_n^{(h)}$	$h_1 = \lvert y-z\rvert$ (Chatterjee)	0.886	0.976	0.994	0.784	0.833	0.837	0.446	0.459	0.462	0.004	-0.003	0.000
		(0.18)	(0.01)	(0.00)	(1.71)	(0.86)	(0.41)	(4.88)	(2.36)	(1.23)	(6.10)	(2.63)	(1.38)
	$h_2 = \lvert y-z\rvert^2$	0.985	0.999	1.000	0.950	0.969	0.971	0.675	0.687	0.691	-0.000	-0.003	0.000
		(0.13)	(0.00)	(0.00)	(0.80)	(0.32)	(0.16)	(5.81)	(2.69)	(1.35)	(9.52)	(4.08)	(2.16)
	$h_3 = \lvert y-z\rvert^3$	0.997	1.000	1.000	0.988	0.994	0.994	0.797	0.808	0.812	-0.008	-0.002	-0.001
		(0.07)	(0.00)	(0.00)	(0.32)	(0.11)	(0.05)	(5.58)	(2.59)	(1.26)	(12.84)	(5.54)	(2.84)
	$h_4 = 1 - e^{\lvert y-z\rvert}$	0.860	0.970	0.992	0.742	0.799	0.803	0.392	0.405	0.408	0.005	-0.003	0.000
		(0.22)	(0.01)	(0.00)	(1.94)	(0.99)	(0.48)	(4.70)	(2.29)	(1.21)	(5.48)	(2.36)	(1.23)
	$h_5 = 1 - e^{(y-z)^2}$	0.982	0.999	1.000	0.941	0.963	0.965	0.638	0.651	0.655	0.002	-0.003	0.000
		(0.15)	(0.00)	(0.00)	(0.93)	(0.38)	(0.18)	(6.01)	(2.76)	(1.41)	(8.65)	(3.71)	(1.98)
	Pearson	-0.380	-0.385	-0.391	-0.381	-0.380	-0.387	-0.311	-0.316	-0.318	0.003	0.007	-0.000
		(7.67)	(3.44)	(1.69)	(7.38)	(3.35)	(1.68)	(8.08)	(3.40)	(1.93)	(8.65)	(4.75)	(2.33)
	Spearman	-0.363	-0.369	-0.376	-0.365	-0.366	-0.375	-0.317	-0.326	-0.326	0.003	0.009	-0.001
		(8.71)	(3.98)	(1.96)	(8.16)	(4.00)	(1.85)	(9.08)	(3.74)	(2.09)	(8.63)	(4.65)	(2.41)
	dCorr	0.440	0.431	0.432	0.442	0.426	0.428	0.371	0.355	0.354	0.160	0.074	0.038
		(4.90)	(2.29)	(1.16)	(4.56)	(2.27)	(1.19)	(5.24)	(2.26)	(1.42)	(3.15)	(1.68)	(0.90)
	MI	0.724	2.002	3.340	0.637	1.430	1.709	0.319	0.508	0.548	0.023	0.011	0.006
		(8.96)	(4.49)	(2.12)	(7.91)	(3.89)	(2.14)	(6.61)	(3.70)	(1.89)	(3.26)	(1.68)	(0.87)
	Theil's U	0.510	0.586	0.638	0.453	0.476	0.452	0.345	0.306	0.265	0.234	0.193	0.159
		(2.80)	(0.93)	(0.34)	(2.73)	(0.90)	(0.41)	(2.64)	(1.00)	(0.45)	(2.14)	(0.79)	(0.29)
	LH	0.269	0.826	0.955	0.262	0.750	0.823	0.128	0.411	0.451	-0.161	-0.032	-0.008
		(2.19)	(0.18)	(0.02)	(2.41)	(0.63)	(0.33)	(3.70)	(1.74)	(0.98)	(2.62)	(0.95)	(0.46)
	MIC	1.000	1.000	1.000	0.990	0.995	0.979	0.671	0.640	0.608	0.238	0.163	0.107
		(0.00)	(0.00)	(0.00)	(1.82)	(0.78)	(0.82)	(8.53)	(3.58)	(2.13)	(3.62)	(1.40)	(0.51)

Statistica Sinica: Newly accepted Paper

selected methods under Model 1 and Model 2. Among these methods, $\xi_n^{(h,F)}$ with $h = h_3$ proves to be the most desirable correlation coefficient under both models, as it exhibits the highest value in scenarios when the noise is small and the dependence is strong, while rapidly approaching zero as X and Y become more independent.

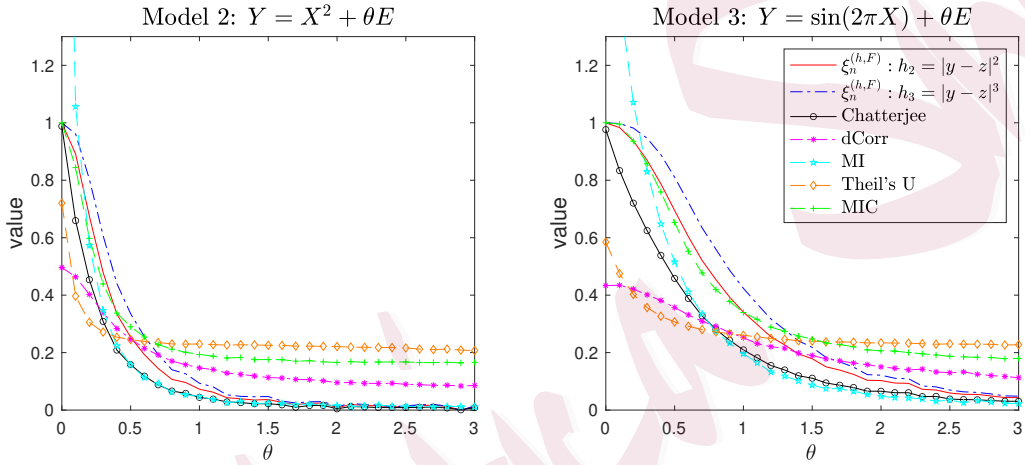

Figure 2: Curves of selected correlation coefficients with varying θ. The sample size is $n = 500$. Experiments are replicated for 100 times.

5.4 Independence tests

This subsection examines the performance of our correlation coefficients in testing the independence hypothesis $H_0 : Y \perp\!\!\!\perp X$. Specifically, we consider the family $\breve{\xi}_n^{(h)}$ with $h(y,z) = |y-z|^\gamma$, where varying $\gamma \in \{0.5, 1, 2, 3\}$. By Theorem 6 and Remark 4.1, the corresponding asymptotic variances

Statistica Sinica: Newly accepted Paper

admit explicit analytical expressions and therefore do not need to be estimated, which makes the implementation of the proposed tests particularly convenient in the simulation study. We consider two additional competing methods: the independence test based on Hoeffding's D, and the power-enhanced version of Chatterjee's correlation coefficient proposed in Lin and Han (2022b) (denoted by "LH," with the number of nearest neighbors set to $M = 5$). The power curves and average p-values under Model 2 are presented in Figure 3.

Several observations can be made from Figure 3. First, for all methods, the Type I error rates under H_0 remain close to the nominal level $\alpha = 0.05$, indicating that all tests provide satisfactory Type I error control. Second, within our proposed family, the power generally increases as γ increases. However, the improvement becomes less pronounced for sufficiently large γ; in particular, the power curves and average p-values for $\gamma = 2$ and $\gamma = 3$ are nearly indistinguishable. Third, the two competing methods exhibit substantially higher power than our proposed methods. This is not entirely surprising, as Shi et al. (2021) theoretically demonstrated that Chatterjee's correlation coefficient is rate-suboptimal under certain local alternatives. Since our proposed family can be viewed as an extension of Chatterjee's correlation coefficient, a similar phenomenon may also arise for our statistics.

Statistica Sinica: Newly accepted Paper

Importantly, Lin and Han (2022b) showed that the power of Chatterjee's correlation coefficient can be substantially improved by increasing the number of nearest neighbors used in its construction, as demonstrated by the "LH" method in Figure 3. This observation suggests a natural direction for improving the power of our proposed family: a similar nearest-neighbor-based power enhancement strategy may potentially be incorporated into our correlation coefficients.

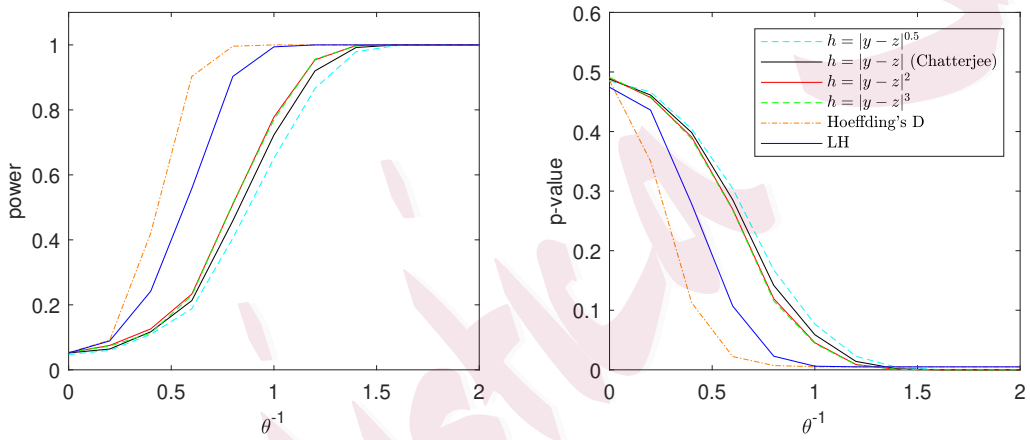

Figure 3: Power curves and average p-values of the selected methods ($\breve{\xi}_n^{(h)}$, Hoeffding's D and LH) under Model 2. Here, $\theta^{-1} = 0$ corresponds to the null hypothesis $H_0 : Y \perp\!\!\!\perp X$, while larger values of θ^{-1} indicate stronger functional dependence. The significance level is $\alpha = 0.05$. The sample size is $n = 1000$, and each experiment is replicated 1000 times.

6. Real data experiment

Statistica Sinica: Newly accepted Paper

6.1 S&P 500 data

The S&P 500, also known as the Standard and Poor's 500, serves as a pivotal stock market index overseeing the performance of approximately 500 of the largest companies listed on United States stock exchanges. A primary objective for financial quant analysts involves examining the longitudinal trends and stability exhibited by these stock prices, which is crucial for refining effective trading strategies. Our proposed correlation coefficient is poised to meet this objective. Specifically, our correlation coefficient is intended for assessing the extent of functional dependence between each stock price Y and the temporal variable X: higher dependence signifies a more stable trend in stock price, while lower dependence corresponds to augmented volatility and unpredictability within the stock. As an illustrative instance, our experiment scrutinizes the short-term volatility of the opening prices of the 504 stocks encompassed by the S&P 500 index over a consecutive 50-day trading span from November 27, 2017, to February 7, 2018. The dataset is accessible at https://www.kaggle.com/datasets/camnugent/sandp500.

Considering that our simplified rank-based correlation coefficient $\breve{\xi}_n^{(h)}$ defined in Definition 3 offers more efficient computation for continuous data, we first apply $\breve{\xi}_n^{(h_2)}$ with $h_2(y,z) = |y-z|^2$ to this real dataset. To assess the effectiveness of our method, we rank the 504 stocks according to their

Statistica Sinica: Newly accepted Paper

corresponding correlation coefficients $\breve{\xi}_n^{(h_2)}$. Figures 4 and 5 present scatter plots of stock prices for the top eight stocks with the highest $\breve{\xi}_n^{(h_2)}$ values and the bottom eight stocks with the lowest $\breve{\xi}_n^{(h_2)}$ values, respectively. In sharp contrast, the eight stocks shown in Figure 4 exhibit relatively smooth trends over time, whereas those in Figure 5 display pronounced volatility. A closer inspection of the top eight stocks in Figure 4 reveals that, apart from IRM—which shows a uniquely monotonic downward trend—the other seven stocks share a similar pattern of first rising and then falling, suggesting a strong interdependence among them. This finding not only demonstrates the capability of our correlation coefficient to identify perfect functional dependence, whether monotonic or non-monotonic, but also highlights its ability to capture groups of stocks with highly similar trajectories, underscoring its practical significance.

To draw a comparison, we also employ five other methods: our $\breve{\xi}_n^{(h)}$ with h function taking $h_3(y,z) = |y-z|^3$, the Chatterjee's correlation coefficient, the MIC (Reshef et al., 2011), the distance correlation (Székely et al., 2007), and the mutual information (estimated by the KSG method (Kraskov et al., 2004)). Table 3 displays the values of each correlation coefficient for the top 8 stocks (selected by $\breve{\xi}_n^{(h_2)}$). The coefficient $\breve{\xi}_n^{(h_3)}$ shows only minor differences from $\breve{\xi}_n^{(h_2)}$, while yielding nearly identical rankings of

Statistica Sinica: Newly accepted Paper

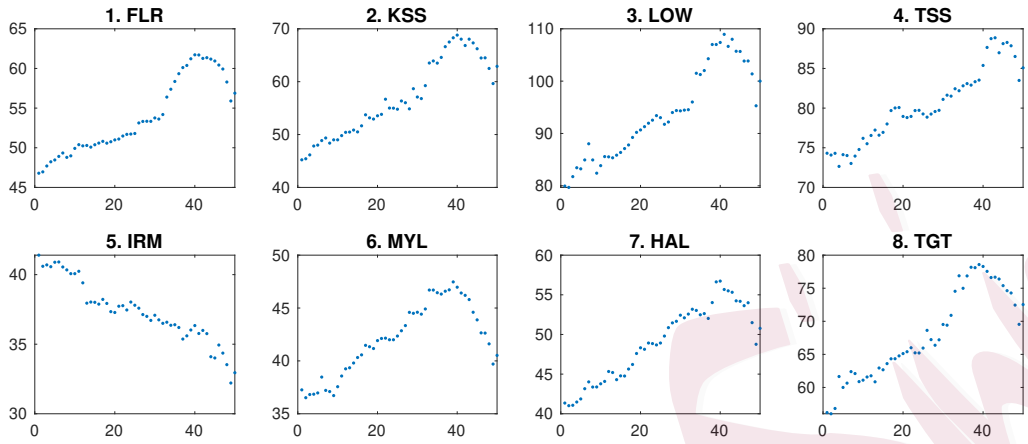

Figure 4: S&P 500 stock data: top 8 stocks picked by $\breve{\xi}_n^{(h_2)}$. In each subplot, the horizontal axis represents the trading days (sequentially numbered from 1 to 50), and the vertical axis represents the stock price (measured in US dollars).

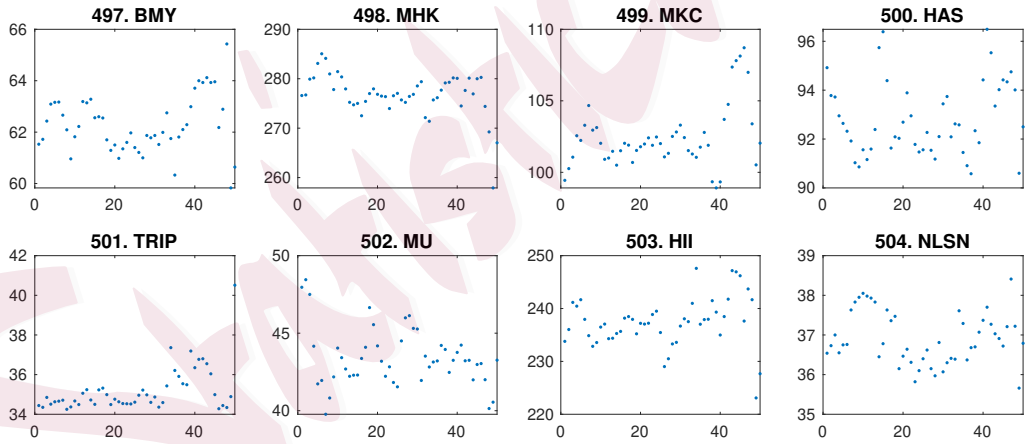

Figure 5: S&P 500 stock data: bottom 8 stocks picked by $\breve{\xi}_n^{(h_2)}$. In each subplot, the horizontal axis represents the trading days (sequentially numbered from 1 to 50), and the vertical axis represents the stock price (measured in US dollars).

Statistica Sinica: Newly accepted Paper

the eight selected stocks among the total 504 stocks. A similar outcome is observed for Chatterjee's correlation coefficient, which in fact corresponds to $\breve{\xi}_n^{(h_1)}$ under our framework of simplified rank-based coefficients. Overall, these three measures belong to the same family, differing only in the choice of the h function. Consequently, their stock rankings are expected to be close, though not exactly identical, which is consistent with our theoretical understanding. If we look beyond the rankings and focus solely on the coefficient values in Table 3, our $\breve{\xi}_n^{(h_2)}$ and $\breve{\xi}_n^{(h_3)}$ are evidently closer to 1 than Chatterjee's coefficient. Given that these eight stocks visually exhibit almost perfect functional dependence, the relatively low values of Chatterjee's coefficient may be somewhat misleading. In contrast, choosing $h_2(y,z) = |y-z|^2$ provides a more faithful characterization of the underlying relationships.

For the MIC, it turns out that 177 out of the total 504 stocks share a MIC value of 1 in common, and 6 out of the 8 selected stocks in Table 3 has MIC of 1. Consequently, the MIC demonstrates limited utility as it fails to discern differences among many of these stocks. The distance correlation appears to be more sensitive in detecting monotonic functional relationships. Among the eight stocks in Figure 4, TSS and IRM exhibit the most monotonic overall trends, and their corresponding dCorr ranks

Statistica Sinica: Newly accepted Paper

are the highest within this group. In contrast, MYL displays a distinctly non-monotonic, inverted-V pattern, and its dCorr rank drops as low as 179. The rankings based on mutual information appear to be the most unstable among all the measures. For instance, stocks such as KSS and LOW receive rankings identical to ours, at second and third place, respectively. However, others such as FLR, IRM, and MYL are ranked 76th, 309th, and 124th, which deviate substantially from the results of the other measures. In particular, IRM exhibits a clear monotonic decreasing functional relationship, yet its MI rank falls as low as 309. This indicates that mutual information is not very sensitive in detecting monotonic functional relationships.

In summary, since there is no ground-truth ranking of the "smoothness" or "strength of functional dependence" among these stock prices, the above analysis cannot be used to directly validate the rankings produced by our method. Nevertheless, the observed patterns provide a reasonable illustration of how the proposed coefficients can be used to characterize functional dependence in real-world data, highlighting their potential practical utility.

Statistica Sinica: Newly accepted Paper

Table 3: Values of selected correlation coefficients for the top 8 chosen stocks, along with their respective ranks among the total of 504 stocks.

Stock	$\check{\xi}_n^{(h_2)}$		$\check{\xi}_n^{(h_3)}$		Chatterjee		MIC		dCorr		MI	
	value	rank	value	rank	value	rank	value	rank	value	rank	value	rank
FLR	0.989	1	0.999	1	0.885	1	1.000	1	0.944	27	0.927	76
KSS	0.979	2	0.997	2	0.854	2	0.971	178	0.954	18	1.295	2
LOW	0.977	3	0.996	3	0.848	3	1.000	1	0.948	21	1.272	3
TSS	0.972	4	0.995	4	0.828	10	0.971	178	0.963	11	1.073	33
IRM	0.971	5	0.995	5	0.830	9	1.000	1	0.957	17	0.447	309
MYL	0.971	6	0.995	6	0.833	5	1.000	1	0.832	179	0.809	124
HAL	0.971	7	0.993	11	0.843	4	1.000	1	0.940	28	1.035	41
TGT	0.970	8	0.994	7	0.833	5	1.000	1	0.939	30	1.176	11

6.2 World birth and death rates

In the second example, we examine the correlation between the birth rate X and death rate Y across 236 countries and territories worldwide in 2023. The data can be accessed at https://ourworldindata.org/grapher/birth-rate-vs-death-rate, with the scatter plots shown in Figure 6. A preliminary look at the data plots reveals a clear dependence between X and Y, though the relationship is neither linear nor monotonic. From the left panel of Figure 6, Y initially exhibits a sharp decreasing trend as X increases. However, after X reaches a turning point approximately at 12, Y begins to increase slightly with further increases in X. This observation suggests some level of functional dependence of Y on X, i.e., $Y = f(X) + \epsilon$ for some non-monotone function f. Note that such functional relationship is in general not symmetric in X

Statistica Sinica: Newly accepted Paper

and Y, unless f is monotonic or invertible. This asymmetry is reflected in the right panel of Figure 6, where, upon transposing the data points, the functional relationship of X on Y is rarely observed.

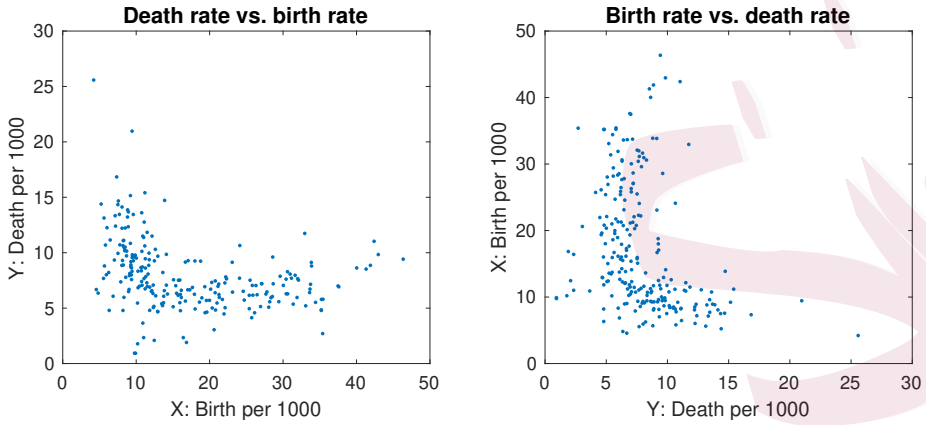

Figure 6: Birth and death rates in 236 countries and territories worldwide in 2023. X: birth per 1000 individuals per country; Y: death per 1000 individuals per country. Left panel: death rate vs. birth rate; right panel: birth rate vs. death rate.

Table 4: Correlation coefficients for the birth and death rate data. The values in parentheses are the corresponding p-values for testing the null hypothesis $H_0 : Y \perp\!\!\!\perp X$.

	$h_1 = \lvert y-z\rvert$	$h_2 = \lvert y-z\rvert^2$	$h_3 = \lvert y-z\rvert^3$	$h_4 = 1-e^{\lvert y-z\rvert}$	$h_5 = 1-e^{(y-z)^2}$
$\xi_n^{(h,F)}(X,Y)$	0.237 ($< 10^{-6}$)	0.350 ($< 10^{-6}$)	0.426 ($< 10^{-6}$)	0.213 ($< 10^{-6}$)	0.328 ($< 10^{-6}$)
$\xi_n^{(h,F)}(Y,X)$	0.104 (0.0018)	0.132 (0.0213)	0.145 (0.1072)	0.097 (0.0023)	0.128 (0.0176)
$\check{\xi}_n^{(h)}(X,Y)$	0.236 ($< 10^{-6}$)	0.344 ($< 10^{-6}$)	0.411 ($< 10^{-6}$)	0.211 ($< 10^{-6}$)	0.323 ($< 10^{-6}$)
$\check{\xi}_n^{(h)}(Y,X)$	0.143 (0.0002)	0.218 (0.0004)	0.266 (0.0010)	0.126 (0.0003)	0.203 (0.0003)

Pearson	Spearman	dCorr	MIC
-0.324	-0.443	0.430	0.427

To assess the sensitivity in capturing such functional dependence of

Statistica Sinica: Newly accepted Paper

Y on X, we apply our proposed correlation coefficients $\xi_n^{(h,F)}(X,Y)$ and $\breve{\xi}_n^{(h)}(X,Y)$ alongside other competitive methods to this dataset. The results are presented in Table 4. For the F function, we adopt $F(y) = \Phi(y-\widehat{\mu}_Y)/\widehat{\sigma}_Y$, where $\widehat{\mu}_Y$ and $\widehat{\sigma}_Y$ denote the mean and the standard deviation respectively of the observed sample Y, and $\Phi(\cdot)$ is the c.d.f. of standard normal distribution. It is seen that $\xi_n^{(h,F)}(X,Y)$ is slightly greater than $\breve{\xi}_n^{(h)}(X,Y)$ when using the same h function; while among all h functions, $h_3 = |y-z|^3$ leads to the greatest value 0.426 of correlation coefficients, demonstrating the highest sensitivity. The Pearson's and Spearman's correlation coefficients result in negative values, which implies some level of negative correlation between X and Y, but are not very informative in capturing functional dependence. The distance correlation "dCorr" yields a value of 0.430, which exceeds all other correlation coefficients. However, from our simulation results in Table 2, it is observed that dCorr ranges between 0.431 and 0.440 under the case of perfect dependence with $\sigma = 0$, indicating that 0.430 is only marginally lower. Consequently, dCorr lacks the specificity needed to differentiate between perfect dependence and noisy dependence, thereby reducing its reliability in this context. The MIC also produces a high value of 0.427. However, as indicated in Table 2, MIC ranges between 0.107 and 0.238 for independent X and Y, which is sig-

Statistica Sinica: Newly accepted Paper

nificantly higher than zero. This indicates that MIC generally produces higher values compared to other correlation coefficients and may exhibit excessive sensitivity. In this context, our proposed correlation coefficients $\xi_n^{(h,F)}(X,Y)$ with h_1, h_2, and h_3 exhibit appropriate sensitivity and provide the more consistent and informative results among the considered choices.

As mentioned, the functional dependence between X and Y may not be symmetric. To give a more comprehensive analysis, we also investigate the functional dependence of X on Y, with results of $\xi_n^{(h,F)}(Y,X)$ and $\breve{\xi}_n^{(h)}(Y,X)$ incorporated in Table 2. Note that in this case, the F function in $\xi_n^{(h,F)}(Y,X)$ need to be switched to $F(x) = \Phi(x - \widehat{\mu}_X)/\widehat{\sigma}_X$, where $\widehat{\mu}_X$ and $\widehat{\sigma}_X$ are the mean and standard deviation of sample X. As expected, all $\xi_n^{(h,F)}(Y,X)$ and $\breve{\xi}_n^{(h)}(Y,X)$ yield significantly lower values than their counterparts $\xi_n^{(h,F)}(X,Y)$ and $\breve{\xi}_n^{(h)}(X,Y)$, which aligns with our initial observations of the data clouds. When the same h function is used, the difference $|\xi_n^{(h,F)}(X,Y) - \xi_n^{(h,F)}(Y,X)|$ is larger than $|\breve{\xi}_n^{(h)}(X,Y) - \breve{\xi}_n^{(h)}(Y,X)|$, and the corresponding difference in the p-values is also larger. These results suggest that $\xi_n^{(h,F)}$ is more sensitive to the dependence asymmetry in this example than the rank-based coefficient $\breve{\xi}_n^{(h)}$. Quantifying this asymmetry in our correlation coefficients is valuable for assessing the extent of non-monotonicity in the dependence of Y on X. This is because, the more

Statistica Sinica: Newly accepted Paper

monotonic the dependence, the closer is $\xi_n^{(h,F)}(X,Y)$ to $\xi_n^{(h,F)}(Y,X)$. This highlights an additional advantage of our proposed correlation coefficients over other candidates, which are strictly symmetric with respect to X and Y and are non-informative in reflecting such dependence asymmetry.

7. Discussion

Inspired by the utility of Chatterjee's correlation coefficient ξ_n in measuring functional dependence, we proposed three new types of correlation coefficients: the original proposal $\xi_n^{(h,F)}$ in Definition 1, the rank-based variant $\widetilde{\xi}_n^{(h)}$ in Definition 2 and its simplified version $\breve{\xi}_n^{(h)}$ in Definition 3. We provide a systematic examination on the asymptotic properties of these coefficients, and demonstrate their capability for measuring functional dependence through both theoretical analysis and case studies.

Several issues warrant further investigation. First, how to select the type of correlation coefficient $\xi_n^{(h,F)}$, $\widetilde{\xi}_n^{(h)}$, or $\breve{\xi}_n^{(h)}$ in a specific real-world problem? From our theoretical results, $\xi_n^{(h,F)}$, $\widetilde{\xi}_n^{(h)}$, and $\breve{\xi}_n^{(h)}$ have similar nice asymptotic properties; while in specific real applications, they may lead to notable differences. Conceptually, $\breve{\xi}_n^{(h)}$ is based on the ranks of Y, suggesting its greater robustness against outliers or abrupt changes; whereas $\xi_n^{(h,F)}$ is non-rank-based, enabling it to leverage more information from the

Statistica Sinica: Newly accepted Paper

distribution of Y sample. These distinct characteristics allow practitioners to select the type of correlation coefficient that best suits their specific requirements. Second, how should one select the functions h and F for our correlation coefficients? Our theoretical results provide useful guidance for this choice. For example, we derive necessary and sufficient conditions on h under which the correlation coefficients possess all the desirable properties (P1)–(P3). Nevertheless, the practical performance of different choices of h and F may vary substantially depending on the intended application. A notable example is the choice of γ within the family $h(y, z) = |y - z|^\gamma$. Our simulation results suggest that, when the primary goal is to detect functional dependence, larger values of γ tend to yield greater sensitivity and faster convergence. In contrast, when the primary goal is independence testing, choosing $\gamma \geq 2$ violates property (P2) and consequently lead to a complete loss of power against certain alternatives. On the other hand, choosing γ too small may result in relatively low power against other types of alternatives, as illustrated by the example in Figure 3. Thus, for independence testing, the choice of γ may need to be tailored to the broad class of alternatives of interest. More generally, there may not be a universally optimal choice of γ across different objectives and evaluation criteria. Developing principled and adaptive strategies for selecting h and F accord-

Statistica Sinica: Newly accepted Paper

ing to the intended purpose and the underlying alternatives is therefore an interesting direction for future research.

Supplementary Materials

The Supplementary Material contains the proofs of main results.

Acknowledgements

We thank the associate editor and the two reviewers for their valuable comments and suggestions. Muhong Gao's work was supported by National Natural Science Foundation of China [Grant No. 12601499], and a start-up fund from University of International Business and Economics. Qizhai Li's work was supported by National Natural Science Foundation of China [Grant No. 12325110, 12288201], the robotic AI-Scientist platform of Chinese Academy of Sciences. Qizhai Li is also affiliated with the School of Mathematical Sciences, University of Chinese Academy of Sciences.

References

Auddy, A., N. Deb, and S. Nandy (2024). Exact detection thresholds and minimax optimality of chatterjee's correlation coefficient. *Bernoulli 30*(2), 1640–1668.

Statistica Sinica: Newly accepted Paper

Azadkia, M. and S. Chatterjee (2021). A simple measure of conditional dependence. *The Annals of Statistics 49*(6), 3070–3102.

Bergsma, W. and A. Dassios (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli 20*(2), 1006–1028.

Bickel, P. J. (2022). Measures of independence and functional dependence. Available at arXiv:2206.13663.

Chang, Y., Y. Li, A. Ding, and J. Dy (2016). A robust-equitable copula dependence measure for feature selection. In *Artificial Intelligence and Statistics*, pp. 84–92. PMLR.

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

Chatterjee, S. (2024). A survey of some recent developments in measures of association. *Probability and Stochastic Processes: A Volume in Honour of Rajeeva L. Karandikar*, 109–128.

Chhaibi, R., F. Gamboa, and C. Pellegrini (2026). A martingale approach to fluctuations of rank estimators in sensitivity analysis. Available at arXiv:2603.22611.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Darbellay, G. and I. Vajda (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory 45*(4), 1315–1321.

Deb, N., P. Ghosal, and B. Sen (2020). Measuring association on topological spaces using kernels

Statistica Sinica: Newly accepted Paper

and geometric graphs. Available at arXiv:2010.01768.

Dette, H., K. F. Siburg, and P. A. Stoimenov (2013). A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics 40*(1), 21–41.

Drton, M., F. Han, and H. Shi (2020). High-dimensional consistent independence testing with maxima of rank correlations. *The Annals of Statistics 48*(6), 3206–3227.

Durrett, R. (2019). *Probability: theory and examples*, Volume 49. Cambridge university press.

Fuchs, S. (2024). Quantifying directed dependence via dimension reduction. *Journal of Multivariate Analysis 201*, 105266.

Fukumizu, K., A. Gretton, B. Schölkopf, and B. K. Sriperumbudur (2008). Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems*, Volume 21. Curran Associates, Inc.

Gamboa, F., P. Gremaud, T. Klein, and A. Lagnoux (2022). Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics. *Bernoulli 28*(4), 2345–2374.

Gao, M., F. Han, and Q. Li (2026). Limit theorems of azadkia-chatterjee's conditional graph correlation. Available at arXiv:2606.15433.

Gao, W., S. Oh, and P. Viswanath (2018). Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory 64*(8), 5629–5661.

Geenens, G. and P. Lafaye de Micheaux (2022). The Hellinger correlation. *Journal of the American Statistical Association 117*(538), 639–653.

Statistica Sinica: Newly accepted Paper

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

Heller, R., Y. Heller, and M. Gorfine (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika 100*(2), 503–510.

Huang, Z., N. Deb, and B. Sen (2022). Kernel partial correlation coefficient—a measure of conditional dependence. *The Journal of Machine Learning Research 23*(1), 9699–9756.

Kinney, J. B. and G. S. Atwal (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences 111*(9), 3354–3359.

Kraskov, A., H. Stögbauer, and P. Grassberger (2004, Jun). Estimating mutual information. *Phys. Rev. E 69*, 066138.

Kvålseth, T. O. (2017). On normalized mutual information: Measure derivations and properties. *Entropy 19*(11).

Li, M., X. Chen, X. Li, B. Ma, and P. Vitanyi (2004). The similarity metric. *IEEE Transactions on Information Theory 50*(12), 3250–3264.

Lin, Z. and F. Han (2022a). Limit theorems of chatterjee's rank correlation. Available at arXiv:2204.08031.

Lin, Z. and F. Han (2022b). On boosting the power of Chatterjee's rank correlation. *Biometrika 110*(2), 283–299.

Statistica Sinica: Newly accepted Paper

Lin, Z. and F. Han (2024). On the failure of the bootstrap for chatterjee's rank correlation. *Biometrika 111*(3), 1063–1070.

Lopez-Paz, D., P. Hennig, and B. Schölkopf (2013). The randomized dependence coefficient. In *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

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

Ma, L. and J. Mao (2019). Fisher exact scanning for dependency. *Journal of the American Statistical Association 114*(525), 245–258.

Nurdiati, S., A. Sopaheluwakan, P. Septiawan, and M. R. Ardhana (2022). Joint spatio-temporal analysis of various wildfire and drought indicators in Indonesia. *Atmosphere 13*(10), 1591.

Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti (2011). Detecting novel associations in large data sets. *Science 334*(6062), 1518–1524.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica 10*(3-4), 441–451.

Sadeghi, B. (2022). Chatterjee Correlation Coefficient: A robust alternative for classic correlation methods in geochemical studies- (including "TripleCpy" Python package). *Ore Geology Reviews 146*, 104954.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society 44*(3), 522–536.

Statistica Sinica: Newly accepted Paper

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

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 2263–2291.

Shi, H., M. Drton, and F. Han (2021). On the power of Chatterjee's rank correlation. *Biometrika 109*(2), 317–333.

Shi, H., M. Hallin, M. Drton, and F. Han (2022). On universally consistent and fully distribution-free rank tests of vector independence. *The Annals of Statistics 50*(4), 1933–1959.

Sriperumbudur, B. K., K. Fukumizu, and G. R. Lanckriet (2011). Universality, characteristic kernels and RKHS embedding of measures. *The Journal of Machine Learning Research 12*(7).

Sudan, M., H. Tyagi, and S. Watanabe (2020). Communication for generating correlation: A unifying survey. *IEEE Transactions on Information Theory 66*(1), 5–37.

Székely, G. J. and M. L. Rizzo (2009). Brownian distance covariance. *The Annals of Applied Statistics 3*(4), 1236–1265.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology 76*(1), 103–154.

Statistica Sinica: Newly accepted Paper

Tishby, N., F. C. Pereira, and W. Bialek (1999). The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377.

Wang, X., W. Pan, W. Hu, Y. Tian, and H. Zhang (2015). Conditional distance correlation. *Journal of the American Statistical Association 110*(512), 1726–1734.

Weihs, L., M. Drton, and N. Meinshausen (2018). Symmetric rank covariances: a generalized framework for nonparametric measures of dependence. *Biometrika 105*(3), 547–562.

Zarei, A. R., M. Mokarram, and M. R. Mahmoudi (2023). Comparison of the capability of the Meteorological and Remote Sensing Drought Indices. *Water Resources Management 37*(2), 769–796.

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

Muhong Gao

School of Statistics, University of International Business and Economics, Beijing 100029, China.

E-mail: gaomh@uibe.edu.cn

Qizhai Li

Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100871, China.

E-mail: liqz@amss.ac.cn