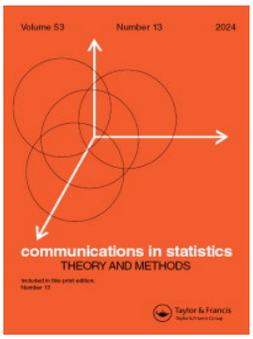

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lsta20

On relationships between Chatterjee's and Spearman's correlation coefficients

Qingyang Zhang

To cite this article: Qingyang Zhang (01 Feb 2024): On relationships between Chatterjee's and Spearman's correlation coefficients, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2024.2309971

To link to this article: https://doi.org/10.1080/03610926.2024.2309971

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2024.2309971

On relationships between Chatterjee's and Spearman's correlation coefficients

Qingyang Zhang

Department of Mathematical Sciences, University of Arkansas, Fayetteville, Arkansas, USA

ABSTRACT

In his seminal work, Chatterjee (2021) introduced a novel correlation measure that is distribution-free, asymptotically normal, and consistent against all alternatives. In this article, we study the probabilistic relationships between Chatterjee's correlation and the widely used Spearman's correlation. We show that, under independence, the two sample-based correlations are asymptotically joint normal and asymptotically independent. Under dependence, the magnitudes of two correlations can be substantially different. We establish some extreme cases featuring large differences between these two correlations. Motivated by these findings, a new independence test is proposed by combining Chatterjee's and Spearman's correlations into a maximal strength measure of variable association. Our simulation study and real-data application show the good sensitivity of the new test to different correlation patterns.

ARTICLE HISTORY
Received 21 February 2023
Accepted 21 January 2024

KEYWORDS
Chatterjee's correlation;
Spearman's correlation;
asymptotic joint normality

1. Introduction

Measuring and testing the dependence between two continuous variables is a durable research topic in statistics. Two classical and arguably the most widely used dependence measures are Pearson's correlation and Spearman's correlation. Pearson's correlation is powerful in detecting linear dependence, especially when the two variables are bivariate normal. Spearman's correlation is a non parametric alternative to Pearson's. It is sensitive to monotonic relations and generally robust to outliers since it is rank-based. Under the null hypothesis of independence, the two sample-based correlations are both asymptotically normal, making it easy to calculate p-values. However, the common drawback of these methods is that they generally fail to detect non monotonic relationships.

In the past decades, there have been numerous tests developed that are consistent against all alternatives, including the kernel-based test Pfister et al. (2018), distance correlation test Székely, Rizzo, and Bakirov (2007), sign covariance test Bergsma and Dassios (2014), copula-based test Schweizer and Wolff (1981), graph-based test Friedman and Rafsky (1983), and maximal information test Reshef et al. (2011), among many others. For a recent survey, see Josse and Holmes (2016). Some of these tests are popular among practitioners, for example, the distance correlation test. However, one major bottleneck of these tests is the testing process: because there is a lack of simple asymptotic theory that facilitates the analytical computation of

CONTACT Qingyang Zhang qz008@uark.edu Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA.

p-values, an expensive permutation test is typically required. For instance, the asymptotic null distribution of distance correlation is difficult to derive because it depends on the underlying distributions of random variables, and the standard approach is to approximate the null distribution of distance covariance via permutation, which requires a time complexity of $O(Rn^2)$, where R is the number of permutations and n is the sample size.

Recently, Chatterjee (2021) introduced a rank-based correlation test that is also consistent with all alternatives Chatterjee (2021). Different from the aforementioned tests, Chatterjee's correlation is asymptotically normal under independence, facilitating quick computation of p-values. Due to its nice properties, Chatterjee's correlation has attracted much attention over the past two years. We begin with a brief review of this method and related literature. Let X and Y be two continuous variables, and $(X_i, Y_i)_{i=1,\ldots,n}$ be n *i.i.d.* samples of (X, Y). Assuming that X_i's and Y_i's have no ties, the data can be uniquely arranged as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$, such that $X_{(1)} < \cdots < X_{(n)}$. Here $Y_{(1)}, \ldots, Y_{(n)}$ denote the concomitants. Let R_i be the rank of $Y_{(i)}$, that is, $R_i = \sum_{k=1}^{n} \mathbb{1}\{Y_{(k)} \le Y_{(i)}\}$, Chatterjee's correlation $\xi_n(X, Y)$ is defined as

$$\xi_n(X, Y) = 1 - \frac{3\sum_{i=1}^{n-1} |R_{i+1} - R_i|}{n^2 - 1}. \tag{1}$$

The asymptotic behavior of $\xi_n(X, Y)$ and related problems have been examined in recent papers. Here we outline a few of them that are most relevant to this work. In his original paper, Chatterjee (2021), Chatterjee showed that $\xi_n(X, Y)$ converges almost surely to the following quantity as n goes to infinity

$$\xi(X, Y) = \frac{\int \mathrm{Var}(E(\mathbb{1}\{Y \ge t|X\}))dF_Y(t)}{\int \mathrm{Var}(\mathbb{1}\{Y \ge t\})dF_Y(t)}.$$

The limiting quantity $\xi(X, Y)$ is also known as Dette-Siburg-Stoimenov's dependence measure, Dette, Siburg, and Stoimenov (2013), which is between 0 and 1 (0 if and only if X and Y are independent, 1 if and only if Y is a measurable function of X). Chatterjee (2021) also established the asymptotic normality of $\xi_n(X, Y)$ under independence. Precisely, $\sqrt{n}\xi_n(X, Y) \xrightarrow{d} N(0, 2/5)$, as $n \to \infty$. The Central Limit Theorem of $\xi_n(X, Y)$ under dependence (as long as Y is not a measurable function of X) is recently proved by Lin and Han (2022). In addition, Auddy, Deb, and Nandy (2021) investigated the limiting power of $\xi_n(X, Y)$ under local alternatives and obtained the exact detection threshold Auddy, Deb, and Nandy (2021). The fast-growing literature on Chatterjee's correlation also includes Shi, Drton, and Han (2022), Shi, Drton, and Han (2021), Lin and Han (2021), Cao and Bickel (2020), Deb, Ghosal, and Sen (2020), Han and Huang (2022), Azadkia and Chatterjee (2021), Zhang (2023), Chatterjee and Vidyasagar (2022), among many others.

In addition to the simple asymptotic theory, as an empirical finding, Chatterjee's test is powerful to detect non monotonic associations, especially those with an oscillating nature such as the W-shaped scatterplot and the sinusoid Chatterjee (2021). The only disadvantage of Chatterjee's test is that it may have less statistical power for smoother alternatives (such as linear or other monotonic relationships) compared to other popular tests, including distance correlation test and Bergsma-Dassios test. For instance, as shown in Figure 5 of Chatterjee (2021), the power of $\xi_n(X, Y)$ quickly deteriorates as the noise level increases, which could be a matter of concern in practice. Motivated by these facts, we propose a versatile test by taking the maximum of Chatterjee's correlation and Spearman's correlation, where the latter

one is powerful to detect monotonic and smoother associations. Two questions arising from this proposal are

(1) What is the asymptotic joint distribution of the two correlations under independence?
(2) How much can they differ as a measure of dependence?

The first question is about the analytical calculation of p-values. The second question investigates if the two correlations to be combined are complementary in the sense that they measure different dependencies. In this article, we give a complete answer to the first question. For the second question, we provide two extreme examples featuring large differences between the two correlations. The idea of combining two complementary correlation metrics is not new. For instance, Zhang, Qi, and Ma (2011) showed the asymptotic independence between Pearson's correlation and a quotient-type correlation and proposed a new test by combining them into a maximal type measure Zhang, Qi, and Ma (2011).

The remainder of this article is structured as follows: Section 2 derives the asymptotic joint distribution of $S_n(X, Y)$ and $\xi_n(X, Y)$ under independence. Section 3 investigates how much the two correlations can differ under dependence. Section 4 proposes the new test of independence, validated by both synthetic data and a real-world dataset. Section 5 discusses the paper with some future perspectives.

2. Asymptotic joint distribution under independence

With the same notations in previous section, Spearman's rank correlation can be written as

$$S_n(X, Y) = 1 - \frac{6\sum_{i=1}^{n}(i - R_i)^2}{n(n^2 - 1)},$$

where R_i represents the rank of $Y_{(i)}, i = 1, ..., n$. Under the hypothesis of independence, it is well known that $E(\sqrt{n}S_n(X, Y)) = 0$, $Var(\sqrt{n}S_n(X, Y)) = n/(n - 1)$, and $\sqrt{n}S_n(X, Y) \xrightarrow{d} N(0, 1)$, as $n \to \infty$. Though $\xi_n(X, Y)$ and $S_n(X, Y)$ are both asymptotically normal, their joint behavior remains unexplored. In this section, we derive the asymptotic joint distribution of $\xi_n(X, Y)$ and $S_n(X, Y)$ under independence.

Let $[n] := \{1, 2, ..., n\}$ be the sample indices. Under independence, $\{R_1, ..., R_n\}$ is a random permutation of $[n]$. We first show that $\xi_n(X, Y)$ and $S_n(X, Y)$ are uncorrelated for a finite sample, as stated in the following lemma:

Lemma 1. *If X and Y are independent, we have*

$$Cov\,[S_n(X, Y), \xi_n(X, Y)] = 0,$$

for any $n \geq 2$.

Proof. Spearman's correlation can be rewritten as

$$S_n(X, Y) = -\frac{3(n + 1)}{n - 1} + \frac{12\sum_{i=1}^{n} iR_i/n}{(n^2 - 1)}.$$

For the covariance between $\xi_n(X, Y)$ and $S_n(X, Y)$, we have

$$\begin{aligned}
&\mathrm{Cov}\left[\sum_{i=1}^{n-1}|R_{i+1}-R_i|, \sum_{j=1}^{n}\frac{j}{n}R_j\right] \\
&= \sum_{i=1}^{n-1}\sum_{j=1}^{n}\mathrm{Cov}\left[R_{i+1}+R_i-2\min(R_{i+1},R_i), \frac{j}{n}R_j\right] \\
&= \sum_{i=1}^{n-1}\sum_{j=1}^{n}\frac{j}{n}\mathrm{Cov}\left[R_{i+1},R_j\right] + \sum_{i=1}^{n-1}\sum_{j=1}^{n}\frac{j}{n}\mathrm{Cov}\left[R_i,R_j\right] \\
&\quad - 2\sum_{i=1}^{n-1}\sum_{j=1}^{n}\frac{j}{n}\mathrm{Cov}\left[\min(R_{i+1},R_i),R_j\right]. \qquad (2)
\end{aligned}$$

The following results (Lin and Han (2021), Lemma 6.1, page 13) are needed for our derivations

$$\begin{aligned}
\mathrm{Cov}[R_1,R_2] &= -\frac{n+1}{12} \\
\mathrm{Var}[R_1] &= \frac{(n-1)(n+1)}{12} \\
\mathrm{Cov}[R_1,\min(R_1,R_2)] &= \frac{(n+1)(n-2)}{24} \\
\mathrm{Cov}[R_1,\min(R_2,R_3)] &= -\frac{n+1}{12}.
\end{aligned}$$

For the first term in Equation (2), we have

$$\begin{aligned}
\sum_{i=1}^{n-1}\sum_{j=1}^{n}\frac{j}{n}\mathrm{Cov}[R_{i+1},R_j] &= \sum_{i=1}^{n-1}\left\{\frac{i+1}{n}\mathrm{Var}[R_1] + \sum_{j\neq i+1}\frac{j}{n}\mathrm{Cov}[R_1,R_2]\right\} \\
&= \sum_{i=1}^{n-1}\left\{\frac{(n+1)(i+1)}{12} - \frac{(n+1)^2}{24}\right\} \\
&= \frac{(n+1)(n-1)}{24}
\end{aligned}$$

Similarly for the second term, we have

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n}\frac{j}{n}\mathrm{Cov}\left[R_i,R_j\right] = -\frac{(n+1)(n-1)}{24}.$$

For the third term, we have

$$\begin{aligned}
&\sum_{i=1}^{n-1}\sum_{j=1}^{n}\frac{j}{n}\mathrm{Cov}\left[\min(R_{i+1},R_i),R_j\right] \\
&= \sum_{i=1}^{n-1}\left\{\frac{2i+1}{n}\mathrm{Cov}[R_1,\min(R_1,R_2)] + \sum_{j\neq i,i+1}\frac{j}{n}\mathrm{Cov}[R_1,\min(R_2,R_3)]\right\}
\end{aligned}$$

Table 1. A special case when $n = 3$.

(R_1, R_2, R_3)	$\xi_3(X, Y)$	$S_3(X, Y)$	$\lvert S_3(X, Y)\rvert$
(1, 2, 3)	1/4	1	1
(1, 3, 2)	−1/8	1/2	1/2
(2, 1, 3)	−1/8	1/2	1/2
(2, 3, 1)	−1/8	−1/2	1/2
(3, 1, 2)	−1/8	−1/2	1/2
(3, 2, 1)	1/4	−1	1

$$= \sum_{i=1}^{n-1} \left\{ \frac{(2i+1)(n+1)(n-2)}{24n} - \frac{(n+1)^2}{24} + \frac{(2i+1)(n+1)}{12n} \right\}$$
$$= 0.$$

Therefore, $\mathrm{Cov}\,[S_n(X, Y), \xi_n(X, Y)] = 0$. This completes the proof of Lemma 1. □

Remark 1. It is noteworthy that Lemma 1 only indicates the uncorrelatedness between $S_n(X, Y)$ and $\xi_n(X, Y)$. In fact, under a finite sample, $S_n(X, Y)$ and $\xi_n(X, Y)$ are generally dependent. A simple example is given in Table 1, where $n = 3$ and $\mathrm{Cov}\,[|S_3(X, Y)|, \xi_3(X, Y)] = 1/24$.

Next, we present a lemma that establishes the Central Limit Theorem for $\{S_n(X, Y), \xi_n(X, Y)\}$. The key steps to prove Lemma 2 include (1) the coupling method for permutation oscillation proposed by Angus (1995) (2) the Central Limit Theorem for m-dependent sequence, and (3) Cramer-Wold device. The detailed proof is a bit lengthy, and we provide it in Appendix.

Lemma 2. *If X and Y are independent, $\sqrt{n}S_n(X, Y)$ and $\sqrt{n}\xi_n(X, Y)$ are asymptotically joint normal.*

By Lemmas 1 and 2, our main theorem follows immediately.

Theorem 1. *If X and Y are independent, $\sqrt{n}S_n(X, Y)$ and $\sqrt{n}\xi_n(X, Y)$ are asymptotically joint normal and asymptotically independent. To be specific,*

$$\begin{bmatrix} \sqrt{n}S_n(X, Y) \\ \sqrt{n}\xi_n(X, Y) \end{bmatrix} \xrightarrow{d} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2/5 \end{pmatrix} \right]$$

as $n \to \infty$.

Theorem 1 answers the first question that we asked in Section 1, which enables the analytical calculation of p-values for the proposed integrated test (to be further discussed in Section 4). Theorem 1, together with Lemma 1 and Remark 1, give a complete characterization for the joint behavior of $S_n(X, Y)$ and $\xi_n(X, Y)$ under independence. The convergence of $\{S_n(X, Y), \xi_n(X, Y)\}$ to joint normality, as sample size n increases, is illustrated in the Figure 1 below.

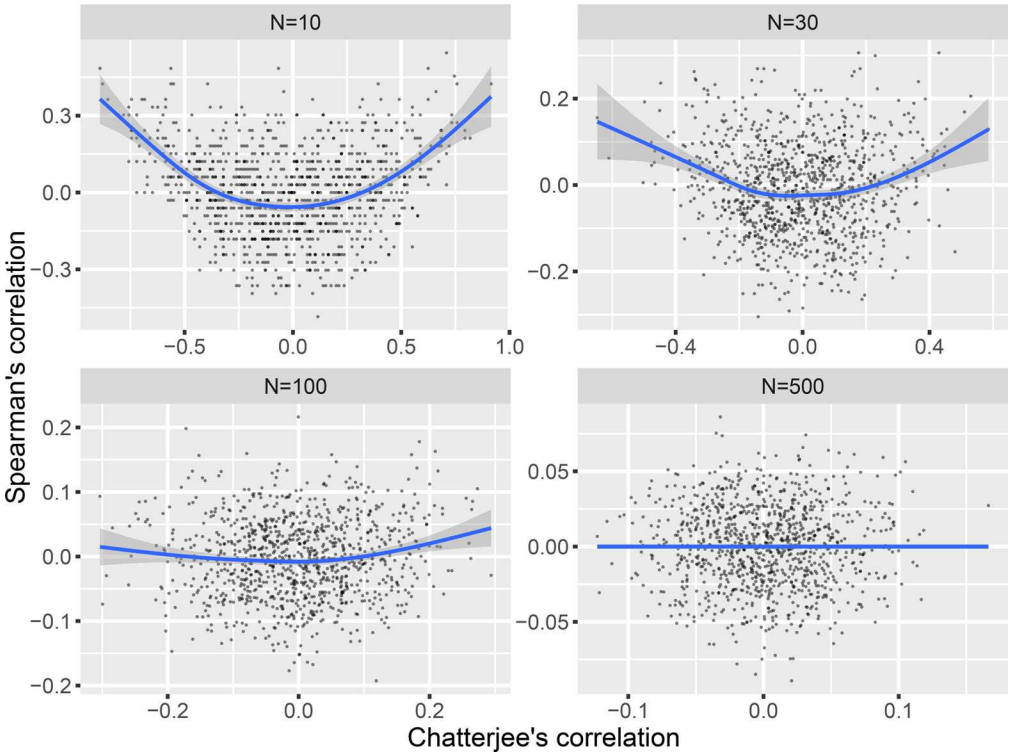

Figure 1. Scatterplots of $S_n(X, Y)$ and $\xi_n(X, Y)$ under $n = 10, 30, 100, 500$.

3. Chatterjee's and Spearman's correlations - how much can they differ?

In this section, we explore the second question outlined in Section 1, that is, how much the two correlations can differ as a metric of dependence. We focus on some extremal cases where the magnitudes of $\xi_n(X, Y)$ and $S_n(X, Y)$ are largely different. For dependent variables X and Y, $\xi_n(X, Y)$ is generally, though not always, between 0 and 1, while $S_n(X, Y)$ is between -1 and 1, therefore we take the absolute value of $S_n(X, Y)$, and compare $\xi_n(X, Y)$ and $|S_n(X, Y)|$ instead. We first provide an extremal case where the absolute Spearman's correlation is small but Chatterjee's correlation is large. This extremal case is easy to construct using simple symmetric patterns such as $Y = |X|$ or $Y = X^2$, $-1 < X < 1$.

Case 1: For any $\epsilon > 0$, there exist ranks $\{R_1, ..., R_n\}$, such that $|S_n(X, Y)| < \epsilon$ and $\xi_n(X, Y) > 1 - \epsilon$.

Proof. Without loss of generality, suppose n is odd. We construct the following ranks

$$R_i = \begin{cases} n - 2(i-1), & 1 \le i \le (n+1)/2 \\ 2i - (n+1), & (n+3)/2 \le i \le n. \end{cases}$$

It is straightforward to show that

$$\xi_n(X, Y) = 1 - \frac{6n - 9}{n^2 - 1}.$$

To derive $S_n(X, Y)$, first we have

$$\begin{aligned}\sum_{i=1}^{n}(i-R_i)^2 &= \sum_{i=1}^{(n+1)/2}(i-R_i)^2 + \sum_{i=(n+3)/2}^{n}(i-R_i)^2 \\ &= \frac{(n-1)(n+1)(n+2)}{8} + \frac{n(n+1)(n-1)}{24} \\ &= \frac{(n-1)(n+1)(2n+3)}{12},\end{aligned}$$

then

$$|S_n(X, Y)| = \frac{3}{2n}.$$

For any given $\epsilon > 0$, we can find an odd number n, such that $(6n-9)/(n^2-1) < \epsilon$ and $3/(2n) < \epsilon$, therefore $|S_n(X, Y)| < \epsilon$ and $\xi_n(X, Y) > 1 - \epsilon$. □

Next, we seek an opposite extremal case where $|S_n(X, Y)|$ is large but $\xi_n(X, Y)$ is small. This extremal case is not straightforward because when $|S_n(X, Y)| = 1$, $\{R_1, ..., R_n\}$ are monotonically increasing or decreasing, therefore $\xi_n(X, Y) = (n-2)/(n+1)$, which means when $|S_n(X, Y)|$ is close to 1, the minimum possible value of $\xi_n(X, Y)$ may not be close to 0. Mathematically, this can be formulated as the following optimization problem

For a given $0 < \epsilon < 1$, find

$$\max_{\{R_1,...,R_n\} \overset{\text{perm}}{=} [n]} \left| 1 - \frac{6\sum_{i=1}^{n}(i-R_i)^2}{n(n^2-1)} \right|,$$

where $\{R_1, ..., R_n\} \overset{\text{perm}}{=} [n]$ represents that $\{R_1, ..., R_n\}$ is a permutation of $\{1, ..., n\}$, given the following inequality constraint

$$1 - \frac{3\sum_{i=1}^{n-1}|R_{i+1} - R_i|}{n^2-1} < \epsilon.$$

Unless n is small enough to enumerate all permutations of $\{R_1, ..., R_n\}$, the optimization problem above is difficult because of the complicated constraints. It may require advanced integer programming techniques, which are beyond the scope of this work. We leave the optimization problem as an open question and try to give a simple example instead, where $\xi_n(X, Y)$ is relatively small but $|S_n(X, Y)|$ is substantially larger. Our intuition is that Spearman's correlation measures the overall monotonic relations, while Chatterjee's correlation is sensitive to local changes. Accordingly, we construct a case that is overall monotonic but has wiggly local patterns.

Case 2: For any $\epsilon > 0$, there exist ranks $\{R_1, ..., R_n\}$, such that $\xi_n(X, Y) = \epsilon + O(1/n)$ and $|S_n(X, Y)| = 1 - \sqrt{2/27}(1-\epsilon)^{3/2} + O(1/n)$.

Proof. We construct $n = 2m + p$ ranks which can be partitioned into two parts: the part of $1 \le i \le 2m$ has an oscillating pattern, while the part of $2m + 1 \le i \le n$ is monotonically increasing

$$R_i = \begin{cases} (i+1)/2, & 1 \le i \le 2m \text{ and } i \text{ is odd} \\ i/2 + m, & 1 \le i \le 2m \text{ and } i \text{ is even} \\ i, & 2m+1 \le i \le n. \end{cases}$$

Let $c = p/m$, the following results can be obtained

$$\begin{aligned}\xi_n(X,Y) &= 1 - \frac{3[m^2 + (m-1)^2 + p]}{(2m+p)^2 - 1} \\ &= 1 - \frac{6}{(c+2)^2} + O(1/n),\end{aligned}$$

$$\begin{aligned}|S_n(X,Y)| &= 1 - \frac{2m(m+1)(2m+1) - 6m^2}{[(2m+p)^2 - 1](2m+p)} \\ &= 1 - \frac{4}{(c+2)^3} + O(1/n).\end{aligned}$$

For any $1 > \epsilon > 0$, there exists $c > 0$ such that $\epsilon = 1 - 6/(c+2)^2$, therefore $\xi_n(X,Y) = \epsilon + O(1/n)$. By the same c, we have $|S_n(X,Y)| = 1 - \sqrt{2/27}(1-\epsilon)^{3/2} + O(1/n)$. □

We give two examples for this extremal case (1) when $n = 100$, $m = 40$ and $p = 20$, $\xi_n(X,Y) \approx 0.058$ while $S_n(X,Y) \approx 0.753$ (2) when $n = 60$, $m = 23$ and $p = 14$, $\xi_n(X,Y) \approx 0.144$ while $S_n(X,Y) \approx 0.789$, both show substantial difference between the two metrics.

Beyond extremal cases 1 and 2, we explore the magnitude of the difference when both coefficients clearly show dependence, that is, when both $|S_n|$ and ξ_n exceed a certain threshold, for example 0.4. Case 3 is constructed such that $|S_n|$ is close to any given threshold η while ξ_n approaches 1.

Case 3: For any $1 > \eta > 0$, there exist ranks $\{R_1, ..., R_n\}$, such that $|S_n(X,Y)| = \eta + O(1/n)$ and $\xi_n(X,Y) = 1 + O(1/n)$.

Proof. Without loss of generality, suppose $n = m + p$, where m is odd. We construct the following ranks

$$R_i = \begin{cases} m - 2(i-1), & 1 \le i \le (m+1)/2 \\ 2i - (m+1), & (m+3)/2 \le i \le m \\ i, & m+1 \le i \le n \end{cases}$$

It is straightforward to show that

$$\xi_n(X,Y) = 1 - \frac{3(n+m-3)}{n^2-1} = 1 + O(1/n),$$

and

$$S_n(X,Y) = 1 - \frac{(m-1)(m+1)(2m+3)}{2n(n-1)(n+1)} = 1 - (m/n)^3 + O(1/n),$$

For a given $1 > \eta > 0$, one can choose n and m, such that $m/n = (1-\eta)^{1/3}$. As an example, for $\eta = 0.4$, $n = 94$ and $m = 79$, we have $S_n(X,Y) = 0.4$ and $\xi_n(X,Y) = 0.94$. □

In Case 4, $\xi_n(X,Y)$ is close to the given threshold η while $|S_n(X,Y)|$ is significantly larger. This can be easily achieved by adapting Case 2. For instance, when $n = 60$, $m = 38$, we obtain $\xi_n(X,Y) = 0.41$ and $S_n(X,Y) = 0.88$. Figure 2 below presents examples for all four cases discussed here.

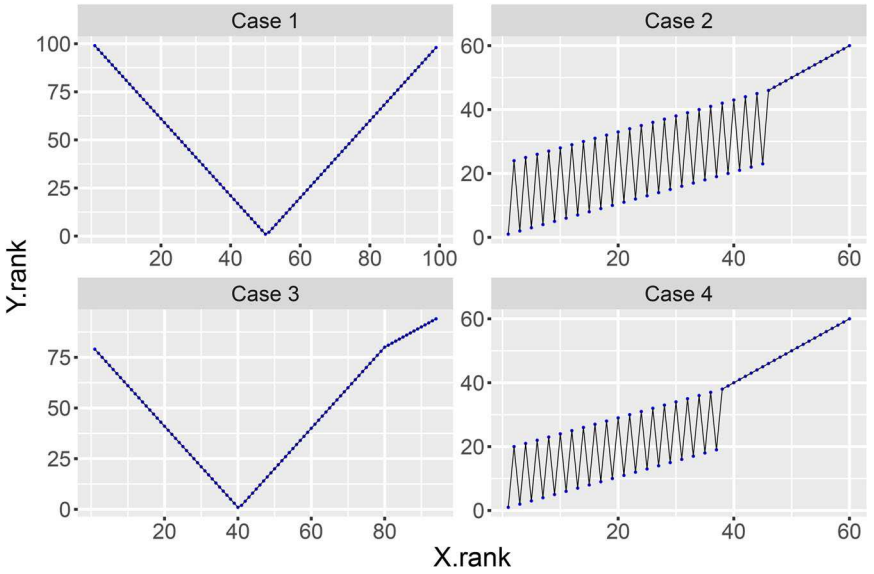

Figure 2. Some illustrative examples for extremal cases 1–4. Case 1: $\xi_n(X,Y) \approx 0.941$, $S_n(X,Y) \approx 0.016$; Case 2: $\xi_n(X,Y) \approx 0.144$, $S_n(X,Y) \approx 0.789$; Case 3: $\xi_n(X,Y) \approx 0.942$, $S_n(X,Y) \approx 0.400$; Case 4: $\xi_n(X,Y) \approx 0.411$, $S_n(X,Y) \approx 0.883$.

4. A new test for independence

Motivated by the findings in Sections 2 and 3, we propose the following new metric

$$I_n(X,Y) = \max\{|S_n(X,Y)|, \sqrt{5/2}\xi_n(X,Y)\}.$$

As $I_n(X,Y)$ takes advantage of both $S_n(X,Y)$ and $\xi_n(X,Y)$, it can be used as a versatile test for detecting both monotonic and non monotonic associations. Moreover, by Theorem 1, one can calculate the asymptotic p-value as follows

$$P(\sqrt{n}I_n(X,Y) > z) \approx 1 - \Phi(z)\left[1 - 2\Phi(-z)\right],$$

where $z \geq 0$ and $\Phi(\cdot)$ represents the standard normal distribution function. For a given significance level of α, we reject the null hypothesis if $\sqrt{n}I_n(X,Y) > c_\alpha$, where c_α satisfies

$$1 - \Phi(c_\alpha)\left[1 - 2\Phi(-c_\alpha)\right] = \alpha.$$

The consistency of the new test under fixed alternatives can be established using Shi, Drton, and Han (2022). Precisely, the testing power satisfies

$$\begin{aligned} P(\sqrt{n}I_n(X,Y) > c_\alpha|H_a) &= P(\max\{\sqrt{n}|S_n(X,Y)|, \sqrt{5n/2}\xi_n(X,Y)\} > c_\alpha|H_a) \\ &\geq P(\sqrt{5n/2}\xi_n(X,Y) > c_\alpha|H_a) \\ &= P(\sqrt{n}\xi_n(X,Y) > \sqrt{2/5}c_\alpha|H_a). \end{aligned}$$

By Proposition 5 of Shi, Drton, and Han (2022), Chatterjee's independence test is consistent for any given α, that is,

$$\lim_{n\to\infty} P(\sqrt{n}\xi_n(X,Y) > \sqrt{2/5}z_{1-\alpha}|H_a) = 1.$$

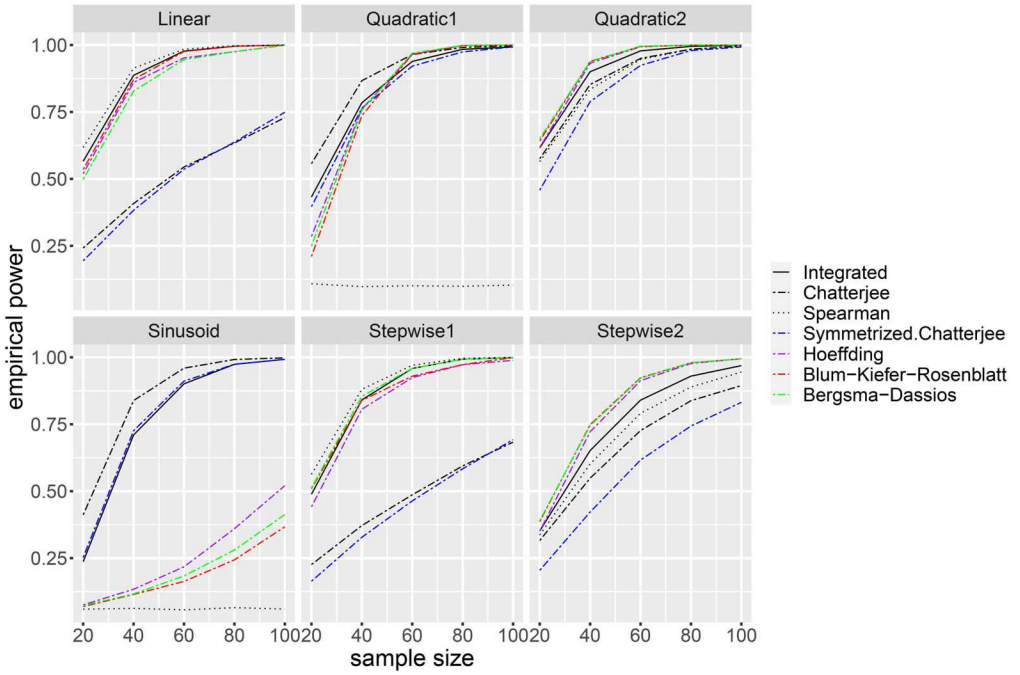

Figure 3. Power comparison of the seven independence tests under different alternatives and sample sizes.

Therefore for any given α, there exists α' such that $c_\alpha = z_{1-\alpha'}$, thus we have

$$\lim_{n\to\infty} P(\sqrt{n} I_n(X,Y) > c_\alpha | H_a) \geq \lim_{n\to\infty} P(\sqrt{n}\xi_n(X,Y) > \sqrt{2/5} c_\alpha | H_a) = 1.$$

We conducted two simulation studies to evaluate the performance of the proposed test. In the first study, we compared the empirical power of seven independence tests, including S_n, ξ_n, I_n, Hoeffding's D Hoeffding (1948), Blum-Kiefer-Rosenblatt's R Blum, Kiefer, and Rosenblatt (1961), Bergsma-Dassios' τ^*Bergsma and Dassios (2014), and the symmetrized version of ξ_n Zhang (2023), under different sample sizes $\{20, 40, 60, 80, 100\}$. The calculations of S_n, ξ_n, and I_n are by our own implementations, and those of D_n, R_n and τ_n^* are made by R package *independence* Even-Zohar (2020). The following six alternatives were considered, where $Z \sim N(0,1)$ and $Z \perp X$

1. Linear: $X \sim Unif(-1,1)$, $Y = X + Z$
2. Quadratic 1: $X \sim Unif(-1,1)$, $Y = X^2 + 0.3Z$.
3. Quadratic 2: $X \sim Unif(-3/4, 5/4)$, $Y = X^2 + 0.4Z$.
4. Sinusoid: $X \sim Unif(-1,1)$, $Y = \cos(2\pi X) + 0.75Z$.
5. Stepwise 1: $X \sim Unif(-1,1)$, $Y = \mathbb{1}_{\{-1\leq X\leq -0.5\}} + 2 * \mathbb{1}_{\{-0.5<X\leq 0\}} + 3 * \mathbb{1}_{\{0<X\leq 0.5\}} + 4 * \mathbb{1}_{\{0.5<X\leq 1\}} + 2Z$.
6. Stepwise 2: $X \sim Unif(-1,1)$, $Y = \mathbb{1}_{\{-1\leq X\leq -0.5\}} + 2 * \mathbb{1}_{\{-0.5<X\leq 0\}} + 3 * \mathbb{1}_{\{0<X\leq 0.5\}} + 2 * \mathbb{1}_{\{0.5<X\leq 1\}} + Z$.

Figure 3 summarizes the empirical power over 5,000 simulation runs (at the significance level of 0.05). As expected, Spearman's test has the highest power for the monotonic settings, that is, "Linear" and "Stepwise 1", but extremely low power for some non monotonic settings including "Sinusoid" and "Quadratic 1". Chatterjee's test is most powerful for two non

Table 2. Empirical size of the seven independence tests ($\alpha = 0.05$).

n	Integrated	Chatterjee	Spearman	Symmetrized	Hoeffding	Blum-Kiefer-Rosenblatt	Bergsma-Dassios
20	0.0406	0.0411	0.0513	0.0225	0.0497	0.0485	0.0490
40	0.0438	0.0462	0.0507	0.0299	0.0503	0.0493	0.0508
60	0.0464	0.0469	0.0499	0.0331	0.0500	0.0501	0.0501
80	0.0489	0.0487	0.0512	0.0333	0.0489	0.0497	0.0492
100	0.0496	0.0492	0.0504	0.0370	0.0502	0.0492	0.0497

monotonic settings, "Quadratic 1" and "Sinusoid", but it has much lower power for the monotonic settings. For instance, in the linear setting when $n = 60$, Chatterjee's test has a power of 0.532, while S_n and I_n both have a power higher than 0.98. The symmetrized ξ_n is slightly less powerful than ξ_n because of its conservativeness, Zhang (2023). Hoeffding's D, Blum-Kiefer-Rosenblatt's R, and Bergsma-Dassios' τ^* have high power in the linear, quadratic, and stepwise settings, but low power in the sinusoidal setting. The new test has satisfactory power for all settings, especially for the monotonic settings where the new test is comparable to Spearman's method. When comparing "Quadratic 2" and "Stepwise 2" settings, where Spearman's and Chatterjee's tests exhibit similar performance, the integrated test demonstrates superior performance, outperforming both. Table 2 summarizes the empirical size over 10,000 simulation runs, where $X \perp Y$, $X \sim \text{Uniform}[-1, 1]$ and $Y \sim N(0, 1)$. It can be seen that all seven tests control the Type I error rate at 0.05. The symmetrized version of ξ_n is slightly conservative.

In the second study, we examined the p-value bias. The exact p-value was approximated using 5,000 permutations and the bias was computed as the asymptotic p-value minus the exact p-value. In each simulation run, we generated X from Uniform$[-1, 1]$ and Y from $N(0, 1)$ independently with sample size $\{20, 40, 60, 80, 100\}$. Figure 4 summarizes the bias over 1,000 simulations runs. It can be seen that the asymptotic p-values are overall close to the exact p-values. However, for a relatively small sample size, for example, $n = 20$, the asymptotic p-values is positively biased, indicating the conservativeness of the test. The bias vanishes as sample size increases. In practice, if the sample size is small, for example, $n < 30$, we recommend a permutation test based on $I_n(X, Y)$ to avoid power loss.

The proposed method was also tested on a transcriptomics dataset by Spellman et al. (1998), which contains the expression levels of 6,223 yeast genes over 23 successive time points during the cell cycle Spellman et al. (1998). This dataset was processed by Reshef et al. (2011), where genes with missing observations were excluded. The processed dataset has 4,381 genes, which are available through the R package *minerva*. There have been many papers testing different correlation measures using this particular dataset including Chatterjee (2021).

We analyzed this data using S_n, ξ_n and I_n. For all three methods, p-values were calculated using asymptotic formulas and then adjusted by the Benjamini-Hochberg procedure to control the false discovery rate (FDR) at the level of 0.05. Figure 5 summarizes the number of significant genes identified by three tests. Out of a total of 4,381 genes, the new test selected 734 genes whose expression levels change during the cell cycle, while the other two tests selected 619 and 385 genes, respectively. This is due to the existence of different expression patterns in the data, that is, some genes have smoother expression change while others have non monotonic such as oscillating expression change. Figure 6 presents a random sample of four genes that were identified by the new test but missed by Spearman's test. It can be seen that the expression levels of all four genes exhibit certain oscillating patterns. Figure 7 shows

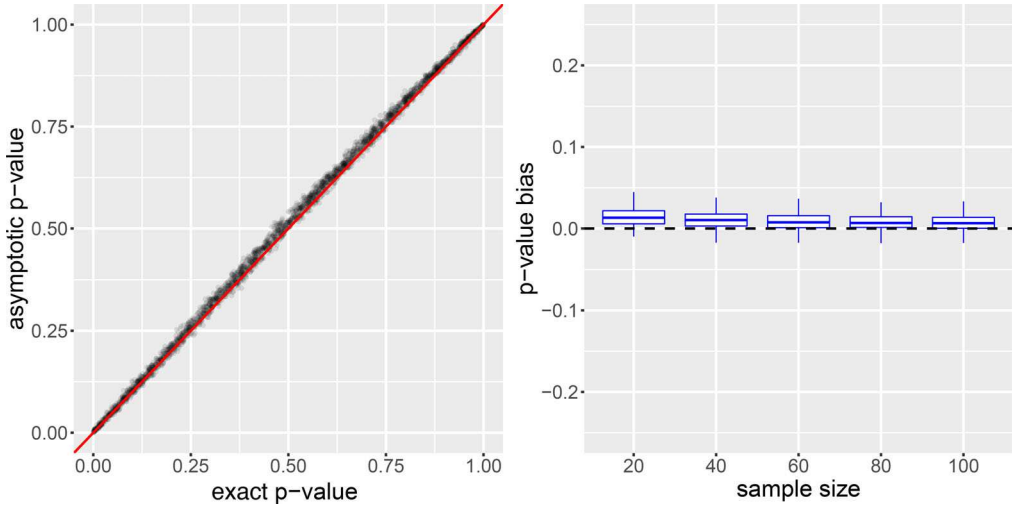

Figure 4. Comparison of the asymptotic and exact p-values. Bias is computed as the asymptotic *p*-value minus the exact *p*-value.

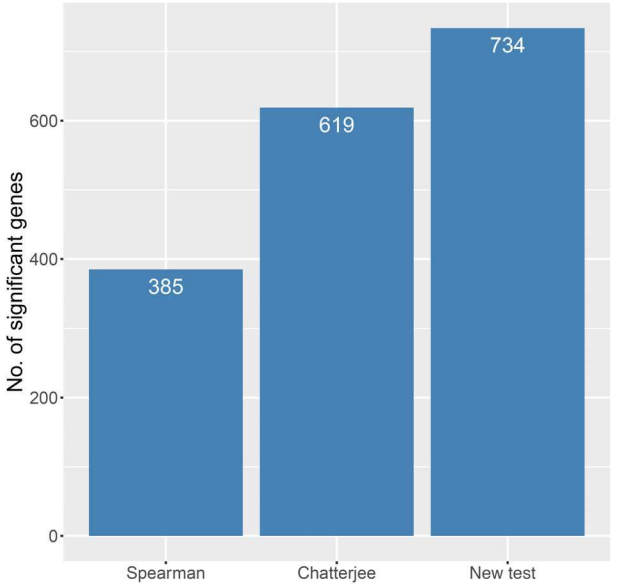

Figure 5. Number of significant genes identified by three methods.

a random sample of four genes that are identified by the new test but missed by Chatterjee's test, where all genes have smoother expression change during the cell cycle.

5. Discussion and conclusions

Chatterjee's rank correlation has attracted a lot of attention during the past two years due to its simplicity and nice statistical properties. However, the cost we pay for this simple method is its inferior performance in detecting smoother correlation patterns, such as linear relationships.

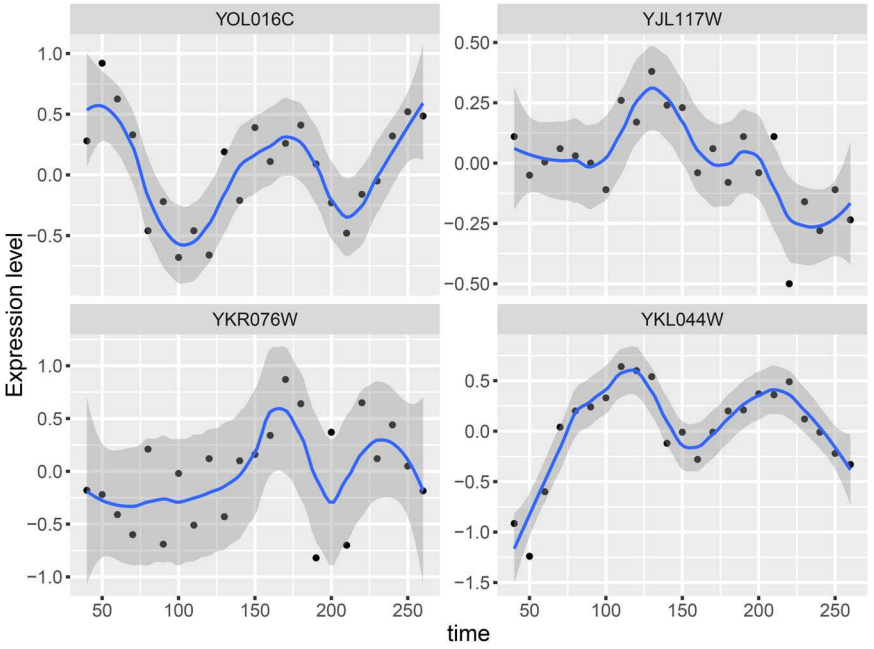

Figure 6. A random sample of four genes selected by the new test but missed by Spearman's test.

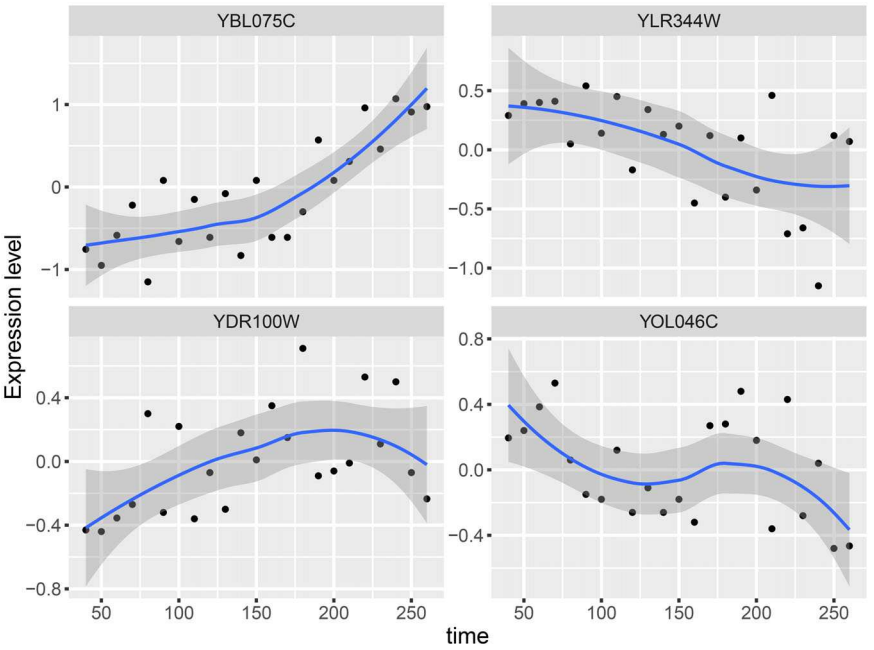

Figure 7. A random sample of four genes selected by the new test but missed by Chatterjee's test.

To boost the power of this ingenious measure, in this article, we proposed a max-type test by combining Chatterjee's correlation with Spearman's correlation, as the latter one is also rank based but sensitive to smooth correlation patterns. We derive the asymptotic joint distribution

of these two correlations under independence, which enables the analytical calculation of p-values. Our simulation study and the transcriptomics application illustrated the promise of the new test. Due to the simple calculation and satisfactory performance, the test is readily applicable to many correlative analyses, for example, the gene–gene interaction and protein–protein interaction network construction.

There are several possible extensions of this work. First, the new test statistic $I_n(X, Y)$ is generally asymmetric because $\xi_n(X, Y)$ is asymmetric, that is, $\xi_n(X, Y) \neq \xi_n(Y, X)$. When a symmetric measure is more suitable, one can consider the following modification

$$I_n^{sym}(X, Y) = \max\{S_n(X, Y), \sqrt{5/2}\xi_n(X, Y), \sqrt{5/2}\xi_n(Y, X)\}.$$

In previous work, Zhang (2023), we established the asymptotic joint normality of $\xi_n(X, Y)$ and $\xi_n(Y, X)$ and showed that the symmetrized metric, that is, $\max\{\xi_n(X, Y), \xi_n(Y, X)\}$, converges to a skew normal distribution under independence. The proof is based on Chatterjee's Central Limit Theorem Chatterjee (2008). The joint asymptotic behavior of $\{S_n(X, Y), \sqrt{5/2}\xi_n(X, Y), \sqrt{5/2}\xi_n(Y, X)\}$ could be studied in a similar way, and the first and most important step is to construct a valid interaction rule for $I_n^{sym}(X, Y)$ Auddy, Deb, and Nandy (2021); Zhang (2023). Although the asymptotic theory of I_n^{sym} remains unknown, it is easy to carry out a permutation test for independence using this statistic. Figure 8 shows the empirical power of I_n^{sym} and I_n under the same simulation settings as those described in Section 4, where I_n^{sym} exhibits satisfactory performance in all settings, albeit with slightly lower power than I_n.

Second, one can consider generalizing our test by replacing ξ_n with its modified version, Lin and Han (2021). The modified statistic $\xi_{n,M}$ incorporates M right nearest neighbors, which is also asymptotically normal but generally more powerful than ξ_n. Notably, Lin and Han (2021) showed that the modified test achieves near-parametric efficiency in testing against Gaussian rotation alternatives. The simulations using permutation tests (Figure 9) confirm that replacing the traditional Chatterjee's coefficient ξ_n with $\xi_{n,M}$ results in consistent power improvement. Notably, the generalized test ($I_{n,M} = \sqrt{n}\max(|S_n|, \sqrt{5M/2}\xi_{n,M})$, where $M = 10, 20$) outperforms the original test across all six settings with different sample sizes, albeit with slight but consistent gains.

Theoretically, it would be important to study the joint behavior of S_n and $\xi_{n,M}$ and derive the asymptotic null distribution of $I_{n,M}$. One possible approach involves leveraging the Hájek representation from Lin and Han (2021). However, a significant obstacle exists concerning the application of Chatterjee's CLT to locally dependent sequences. Recall that $\xi_{n,M}$ is defined as

$$\xi_{n,M} = -2 + \frac{6\sum_{i=1}^{n}\sum_{m=1}^{M}\min\{R_i, R_{m(i)}\}}{(n+1)[nM + M(M+1)/4]},$$

where $m(i)$ is the index of the mth right nearest neighbor of X_i. Lin and Han (2021) established its asymptotic normality under independence by leveraging the Hájek representation of $\xi_{n,M}$ (see $\hat{\xi}_{n,M}$ in Section A.2.1, proof of Theorem 3.2). Their proof hinges on demonstrating the asymptotic equivalence in distribution between $\xi_{n,M}$ and its Hájek representation. This allows them to focus on the asymptotic normality of $\hat{\xi}_{n,M}$, which can be further established using Chatterjee's CLT based on a well-defined interaction rule.

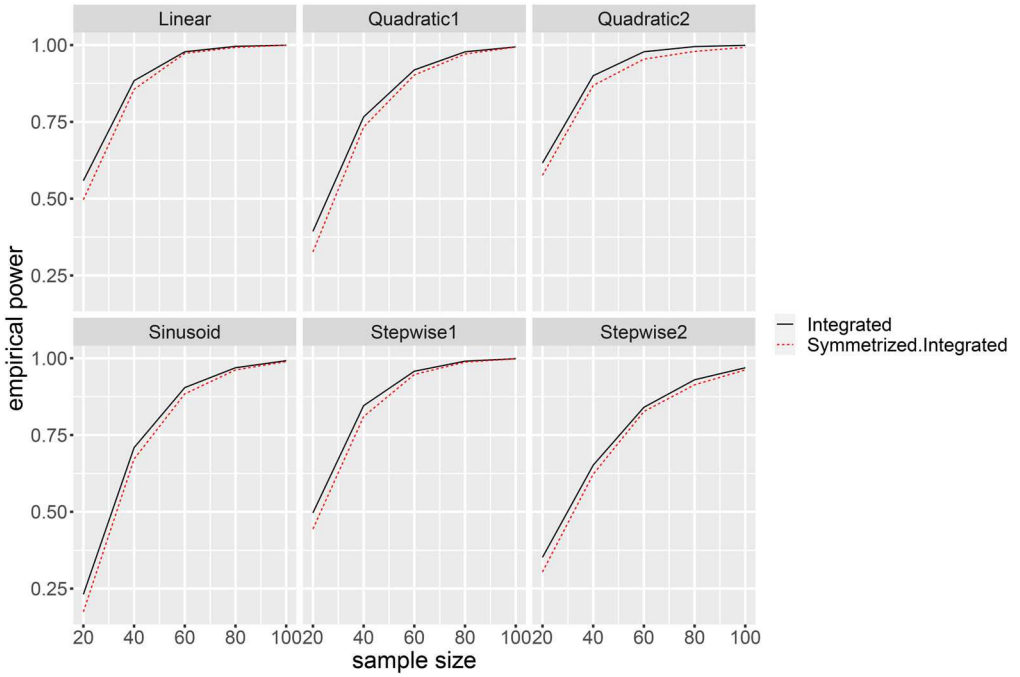

Figure 8. Power comparison of the integrated test and its symmetrized version under different alternatives and sample sizes.

Following this approach, we define the Hájek representation of $a\xi_{n,M} + bS_n$ for constants a and b as follows

$$a\hat{\xi}_{n,M} + b\hat{S}_n = \frac{6an}{(n+1)[nM + M(M+1)/4]} \sum_{i=1}^{n}\sum_{m=1}^{M} \min\{F(Y_i), F(Y_{m(i)})\}$$
$$- \frac{6a}{(n-1)(n+1)} \sum_{i\neq j} \min\{F(Y_i), F(Y_j)\} + b\sum_{i=1}^{n}\left[\frac{2i}{n+1} - 1\right] F(Y_i).$$

Then an asymptotically equivalent version with local dependence is

$$a\check{\xi}_{n,M} + b\check{S}_n = \frac{6an}{(n+1)[nM + M(M+1)/4]} \sum_{i=1}^{n}\sum_{m=1}^{M} \min\{F(Y_i), F(Y_{m(i)})\}$$
$$- \frac{12a}{n+1}\sum_{i=1}^{n}\left(F(Y_i) - \frac{1}{2}F^2(Y_i) - \frac{1}{3}\right) - b\sum_{i=1}^{n}\left[\frac{2i}{n+1} - 1\right] F(Y_i) - \frac{2an}{n+1}.$$

It suffices to show the asymptotic normality of $a\check{\xi}_{n,M} + b\check{S}_n$. However, applying Chatterjee's CLT necessitates a valid interaction rule G that satisfies all conditions outlined in Theorem 2.5 of Chatterjee (2008). The interaction rule employed by Lin and Han (2021), unfortunately, does not apply in this regard. Specifically, the function $f_i(\mathbf{Z})$ in their proof (see Section A.2.1, Step III-1, Lin and Han (2021)) becomes

$$f_i(\mathbf{Z}) = \frac{6an}{(n+1)[nM + M(M+1)/4]} \sum_{m=1}^{M} \min\{F(Y_i), F(Y_{m(i)})\} + \frac{2bi}{n+1}F(Y_i),$$

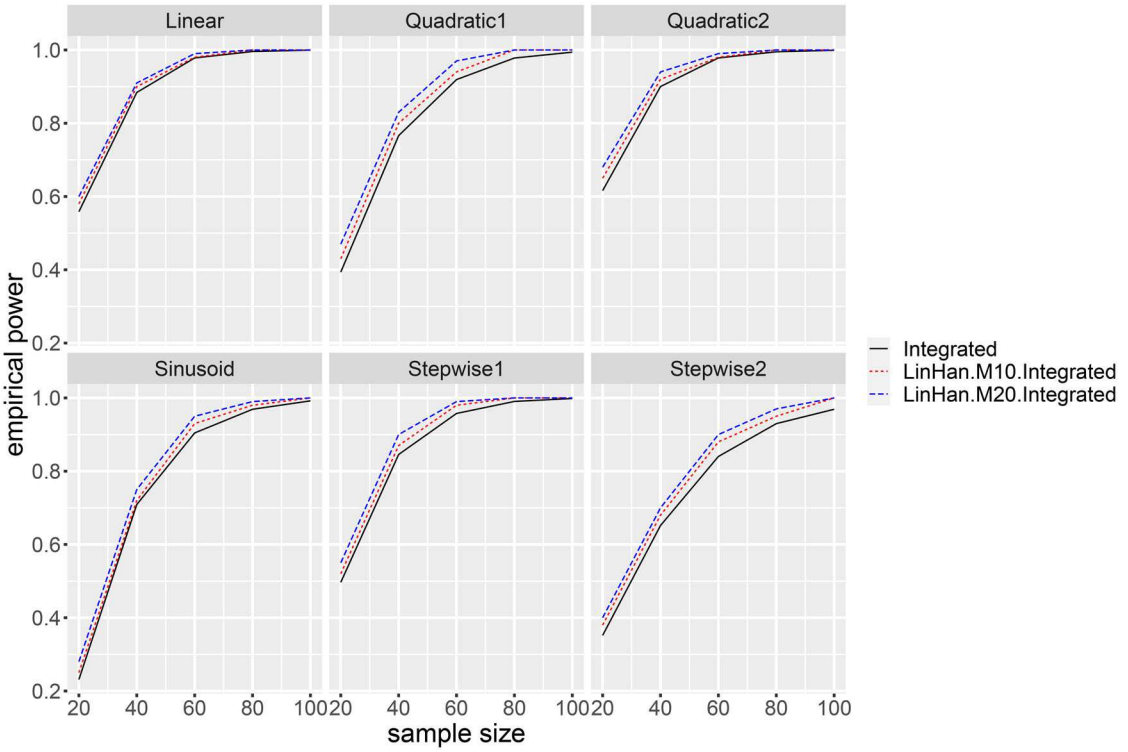

Figure 9. Power comparison of the integrated test and two modifications based on $\xi_{n,M=10}$ and $\xi_{n,M=20}$ under different alternatives and sample sizes.

which in general does not satisfy $f_i(\mathbf{Z}) - f_i(\mathbf{Z}^k) = f_i(\mathbf{Z}^l) - f_i(\mathbf{Z}^{kl})$. Identifying an alternative interaction rule that fulfills all conditions for applying Chatterjee's CLT within our framework remains a critical task for future research.

Third, as enlightened by a reviewer, it is also possible to define the appropriate version of the new test statistic for testing conditional independence, that is, $H_0 : X \perp Y|Z$, although there are some technical hurdles in (1) defining the rank-based estimator of conditional Spearman's correlation and (2) deriving the asymptotic distribution of the integrated statistics. To make it suitable for conditional independence test, first, we can replace $\xi_n(X, Y)$ with the following $T_n(X, Y|Z)$, a multivariate analogue of Chatterjee's correlation, proposed by Azadkia and Chatterjee (2021). Let $(X_i, Y_i, Z_i)_{i=1,\ldots,n}$ be i.i.d. samples of (X, Y, Z)

$$T_n(X, Y|Z) = \frac{\sum_{i=1}^{n}(\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\})}{\sum_{i=1}^{n}(R_i - \min\{R_i, R_{N(i)}\})},$$

where R_i is the rank of Y_i, $M(i)$ is the j such that X_j is the nearest neighbor of X_i, $N(i)$ is the j such that Z_j is the nearest neighbor of Z_i. Azadkia and Chatterjee (2021) showed that $T_n(X, Y|Z)$ converges almost surely to a limit quantity $T(X, Y|Z)$, such that $T(X, Y|Z) = 0$ if and only if X and Y are conditionally independent given Z, and $T(X, Y|Z) = 1$ if and only if Y is almost surely equal to a measurable function of X given Z. Second, we need to replace $S_n(X, Y)$ with its conditional counterpart. However, this presents a challenge due to the absence of a straightforward, rank-based estimator for conditional Spearman's correlation.

The partial rank correlation, defined as

$$S_n(X, Y|Z) = \frac{S_n(X, Y) - S_n(X, Z)S_n(Y, Z)}{\sqrt{(1 - S_n^2(X, Z))(1 - S_n^2(Y, Z))}}$$

is an adaptation of Pearson's partial correlation that replaces the actual values with their ranks. However, as noted in Kendall (1942), this definition lacks justification, and its interpretation is unclear (e.g., whether it truly measures conditional linear relationships). Liu et al. (2018) proposed a sound definition for conditional Spearman's correlation based on a kernel-based estimator involving residual calculations (see Sections 3.2–3.3 in Liu et al. (2018)). This definition, however, is not rank-based, making it challenging to analyze its joint behavior with $T_n(X, Y|Z)$.

Deriving the asymptotic distribution of the integrated statistic presents another significant challenge. Even for the simple rank-based estimator $S_n(X, Y|Z)$, obtaining its asymptotic null distribution is difficult (although simulations suggest normality). For instance, applying Lin and Han (2021)'s method proved unfeasible due to the difficulty in establishing a Hájek representation of $S_n(X, Y|Z)$ that readily transforms into a locally dependent sequence. The Hájek representation of $T_n(X, Y|Z)$ is more straightforward, which can be written as a linear combination (see the representation $\xi_n^{\#}$ in A.2.2, proof of Theorem 3.1, Shi, Drton, and Han (2021)). Additionally, even under conditional independence, the asymptotic distributions of $S_n^2(X, Y)$, $S_n^2(X, Z)$ and $S_n^2(Y, Z)$ remain elusive due to potential pairwise dependencies among X, Y and Z.

Finally, as we discussed in the simulation study, the asymptotic p-value is generally close to the true p-value, but it tends to be positively biased for small sample, for example, $n < 30$, resulting in certain power loss. In the case of small sample, we recommend a permutation test for better testing performance. Another way to reduce the potential p-value bias is to use asymptotic expansion method, for example, Edgeworth expansion, Cornish-Fisher expansion, or saddle point approximation, which may improve p-value approximation by incorporating higher-order moments such as skewness and kurtosis.

Appendix: Proof of Lemma 2

Proof. We first define $F(y) = P(Y < y)$, $U_i = F(Y_{(i)})$, $F_n(y) = \sum_{i=1}^n \mathbb{1}\{Y_{(i)} \le y\}/n$, and $H_n(x) = \sum_{i=1}^n \mathbb{1}\{U_i \le x\}/n$. For Chatterjee's correlation, using Equations (5)–(8) in Angus (1995), we have

$$\frac{\sum_{i=1}^{n-1} |R_{i+1} - R_i| - n(n-1)/3}{\sqrt{n}(n-1)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-1} \left[|U_{i+1} - U_i| + 2U_i(1 - U_i) - \frac{2}{3} \right] + Z,$$

where $Z \xrightarrow{P} 0$. For Spearman's correlation, we define the following function

$$G_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{2i}{n+1} \mathbb{1}\{U_i \le x\}.$$

Since

$$\frac{1}{n} \sum_{i=1}^{n} \frac{2i}{n+1} \mathbb{1}\{U_i \le x\} \le \frac{1}{n} \sum_{i=1}^{n} \frac{2i}{n+1} = 1,$$

we have $0 \leq G_n(x) \leq 1$. The expectation and variance of $G_n(x)$ are $E[G_n(x)] = x$ and

$$\text{Var}[G_n(x)] = \frac{2x(1-x)(2n+1)}{3n(n+1)} \leq \frac{2n+1}{6n(n+1)} \to 0,$$

therefore $G_n(x) \xrightarrow{P} x$ for $x \in [0,1]$, as $n \to \infty$. It is also noteworthy that

$$\begin{aligned}\frac{1}{n\sqrt{n}}\left(\sum_{i=1}^{n}\frac{2i}{n+1}R_i - \frac{n^2}{2}\right) &= \int \sqrt{n}\left[H_n(x) - \frac{1}{2}\right] dG_n(x) \\ &= \int \sqrt{n}\left[H_n(x) - x\right] dG_n(x) + \int \sqrt{n}\left[x - \frac{1}{2}\right] dG_n(x) \quad \text{(A.1)}\end{aligned}$$

where the second term can be rewritten as

$$\int \sqrt{n}\left[x - \frac{1}{2}\right] dG_n(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{2i}{n+1}U_i - \frac{\sqrt{n}}{2}.$$

The first term in Equation (A.1), using continuous mapping theorem, has the same limiting distribution as

$$\int \sqrt{n}\left[H_n(x) - x\right] dx = \frac{\sqrt{n}}{2} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i,$$

therefore

$$\frac{1}{n\sqrt{n}}\left(\sum_{i=1}^{n}\frac{2i}{n+1}R_i - \frac{n^2}{2}\right) \overset{d}{\approx} \sum_{i=1}^{n}\left[\frac{2i}{n+1} - 1\right] U_i.$$

where $\overset{d}{\approx}$ represents asymptotic equivalence in distribution. We will show that $\sum_{i=1}^{n-1}\left[|U_{i+1} - U_i| + 2U_i(1-U_i) - 2/3\right]/\sqrt{n}$ and $\sum_{i=1}^{n}\left[2i/(n+1) - 1\right] U_i/\sqrt{n}$ are asymptotically joint normal.

For any two constants, a and b, define

$$Z_i = a|U_{i+1} - U_i| + 2aU_i(C_i - U_i) - \frac{2a}{3}$$

and

$$W_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} Z_i,$$

where

$$C_i = 1 + \frac{bi}{a(n+1)} - \frac{b}{2a}.$$

It can be seen that for any $j \geq 1$, Z_{i+j} is independent of $[Z_1, ..., Z_i]$, therefore the sequence $\{Z_i\}$ is 1-dependent sequence. Similar to Equations (11)–(14) in Angus (1995), we have

$$\text{Var}[Z_i] = \text{Var}[a|U_{i+1} - U_i|] + \text{Var}[2aU_i(C_i - U_i)] + 2\text{Cov}[a|U_{i+1} - U_i|, 2aU_i(C_i - U_i)],$$

where

$$\text{Var}[a|U_{i+1} - U_i|] = \frac{a^2}{18},$$

and

$$\mathrm{Var}[2aU_i(C_i - U_i)] = 4a^2\left[\frac{C_i^2}{12} - \frac{C_i}{6} + \frac{4}{45}\right].$$

For the covariance term, it can be shown that $\mathrm{Cov}[|U_{i+1} - U_i|, U_i] = 0$, therefore

$$\begin{aligned} 2\mathrm{Cov}[a|U_{i+1} - U_i|, 2aU_i(C_i - U_i)] &= -4a^2\mathrm{Cov}[|U_{i+1} - U_i|, U_i^2] \\ &= -\frac{a^2}{45}, \end{aligned}$$

Summarizing the results above, we get

$$\begin{aligned} \mathrm{Var}[Z_i] &= \frac{a^2}{18} + 4a^2\left[\frac{C_i^2}{12} - \frac{C_i}{6} + \frac{4}{45}\right] - \frac{a^2}{45} \\ &= \frac{a^2}{18} + \frac{b^2}{12}\frac{(2i - n - 1)^2}{(n+1)^2} \\ &\geq \frac{a^2}{18} \end{aligned}$$

For the covariance between Z_i and Z_{i+1}, we have

$$\begin{aligned} \mathrm{Cov}[Z_i, Z_{i+1}] &= \mathrm{Cov}[a|U_{i+1} - U_i| + 2aU_i(C_i - U_i), \\ &\qquad a|U_{i+2} - U_{i+1}| + 2aU_{i+1}(C_{i+1} - U_{i+1})] \\ &= \mathrm{Cov}[a|U_{i+1} - U_i|, a|U_{i+2} - U_{i+1}|] \\ &\quad + \mathrm{Cov}[2aU_i(C_i - U_i), 2aU_{i+1}(C_{i+1} - U_{i+1})] \\ &= \frac{a^2}{180} - \frac{a^2}{90} \\ &= -\frac{a^2}{180}, \end{aligned}$$

therefore

$$\begin{aligned} \mathrm{Var}\left[\sum_{i=1}^{n-1} Z_i\right] &= \sum_{i=1}^{n-1}\mathrm{Var}(Z_i) + 2\sum_{i=1}^{n-1}\mathrm{Cov}(Z_i, Z_{i+1}) \\ &\geq \frac{2(n-1)a^2}{45}, \end{aligned}$$

and

$$\frac{\sqrt{\mathrm{Var}\left[\sum_{i=1}^{n-1} Z_i\right]}}{n^{1/3}} \to \infty$$

as $n \to \infty$. Using the Central Limit Theorem for m-dependent random variables (Theorem 1s in Angus (1995)), W_n converges in distribution to a normal distribution. Finally, by Cramer-Wold device, $\sqrt{n}S_n(X, Y)$ and $\sqrt{n}\xi_n(X, Y)$ are asymptotically joint normal. □

Disclosure statement

The author has declared that no competing interests exist.

Funding

The work was supported by an NSF DBI Biology Integration Institute (BII) grant (award no. 2119968; PI-Ceballos).

References

Angus, J. E. 1995. A coupling proof of the asymptotic normality of the permutation oscillation. *Probability in the Engineering and Informational Sciences* 9 (4):615–21. doi:10.1017/S0269964800004095.

Auddy, A., N. Deb, and S. Nandy. 2021. Exact detection thresholds and minimax optimality of Chatterjee's correlation coefficient. arXiv preprint. arXiv:2104.15140.

Azadkia, M., and S. Chatterjee. 2021. A simple measure of conditional dependence. *Annals of Statistics* 49 (6):3070–102.

Bergsma, W., and A. Dassios. 2014. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli* 20 (2):1006–28.

Blum, J. R., J. Kiefer, and M. Rosenblatt. 1961. Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics* 32 (2):485–98. doi:10.1214/aoms/1177705055.

Cao, S., and P. Bickel. 2020. Correlations with tailored extremal properties. arXiv preprint. arXiv:2008.10177.

Chatterjee, S. 2008. A new method of normal approximation. *Annals of Probability* 36 (4):1584–610.

Chatterjee, S. 2021. A new coefficient of correlation. *Journal of the American Statistical Association* 116 (536):2009–22. doi:10.1080/01621459.2020.1758115.

Chatterjee, S., and M. Vidyasagar. 2022. Estimating large causal polytree skeletons from small samples. arXiv preprint. arXiv:2209.07028.

Deb, N., P. Ghosal, and B. Sen. 2020. Measuring association on topological spaces using kernels and geometric graphs. arXiv preprint. 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. doi:10.1111/j.1467-9469.2011.00767.x.

Even-Zohar, C. 2020. *independence: Fast rank-based independence testing*, R package version 1.0.1, https://CRAN.R-project.org/package=independence.

Friedman, J. H, and L. C. Rafsky. 1983. Graph-theoretic measures of multivariate association and prediction. *Annals of Statistics* 11 (2):377–91.

Han, F., and Z. Huang, 2022. Azadkia-Chatterjee's correlation coefficient adapts to manifold data. arXiv preprint. arXiv:2209.11156.

Hoeffding, W. 1948. A non-parametric test of independence. *The Annals of Mathematical Statistics* 19 (4):546–57. doi:10.1214/aoms/1177730150.

Josse, J., and S. Holmes. 2016. Measuring multivariate association and beyond. *Statistics Surveys* 10:132–67. doi:10.1214/16-SS116.29081877.

Kendall, M. G. 1942. Partial rank correlation. *Biometrika* 32 (3-4):277–83. doi:10.1093/biomet/32.3-4.277.

Lin, Z., and F. Han. 2021. On boosting the power of Chatterjee's rank correlation. *Biometrika* 110 (2):283–299. doi:10.1093/biomet/asac048.

Lin, Z., and F. Han. 2022. Limit theorems of Chatterjee's rank correlation. arXiv preprint. arXiv:2204.08031.

Liu, Qi., C. Li, V. Wanga, and B. E. Shepherd. 2018. Covariate-adjusted Spearman's rank correlation with probability-scale residuals. *Biometrics* 74 (2):595–605. doi:10.1111/biom.12812.29131931.

Pfister, N., P. Bühlmann, B. Schölkopf, and J. Peters. 2018. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80 (1):5–31. doi:10.1111/rssb.12235.

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 (New York, N.Y.)* 334 (6062):1518–24. doi:10.1126/science.1205438. 22174245.

Schweizer, N., and E. F. Wolff. 1981. On nonparametric measures of dependence for random variables. *Annals of Statistics* 9 (4):879–85.

Shi, H., M. Drton, and F. Han. 2021. On Azadkia-Chatterjee's conditional dependence coefficient. arXiv preprint. arXiv:2108.06827.

Shi, H., M. Drton, and F. Han. 2022. On the power of Chatterjee's rank correlation. *Biometrika* 109 (2):317–33. doi:10.1093/biomet/asab028.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* 9 (12):3273–97. doi:10.1091/mbc.9.12.3273.9843569.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *Annals of Statistics* 35 (6):2769–94.

Zhang, Q. 2023. On the asymptotic null distribution of the symmetrized Chatterjee's correlation coefficient. *Statistics & Probability Letters* 194:1–7. doi:10.1016/j.spl.2022.109759.

Zhang, Z., Y. Qi, and X. Ma. 2011. Asymptotic independence of correlation coefficients with application to testing hypothesis of independence. *Electronic Journal of Statistics* 5:342–72.