On the Metricity of the Chatterjee Correlation Coefficient*

Flavio Chierichetti†
Sapienza University of Rome
and
Mirko Giacchini†
Sapienza University of Rome
and
Ravi Kumar†
Google, Mountain View, CA

February 21, 2026

Abstract

We show that the distance measure implied by the recently proposed Chatterjee coefficient of correlation can violate the triangle inequality, both in theory and in practice.

Keywords: Triangle inequality, Transitivity, Distance measure

*This is an Accepted Manuscript of an article published by Taylor & Francis in The American Statistician on 19 Nov 2025, available at: https://doi.org/doi:10.1080/00031305.2025.2571183

†Correspondence to: flavio@di.uniroma1.it, giacchini@di.uniroma1.it, ravi.k53@gmail.com

1 Introduction

Correlation measures are central objects of study in statistics. They are designed to quantify the relationships between variables, which is essential to understand latent patterns in data. Traditional correlation measures include the Pearson and Spearman coefficients. The former quantifies linear relationship between variables whereas the latter quantifies monotonic relationships. Correlation measures are ubiquituous and find applications in various fields including econometrics, social sciences, machine learning, etc.

Recently Chatterjee (2021) proposed a simple estimator of a correlation coefficient between two random variables that has an asymptotic theory under the hypothesis of independence (Lin and Han (2022a) extended this result beyond the independent case), it is non-parametric, robust to outliers, and performs well in non-linear and non-monotonic settings, thus offering many advantages over Pearson and Spearman coefficients.

Let (X, Y) be a pair of random variables such that Y is non-constant, and let the pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent samples from (X, Y). Without loss of generality, suppose $X_1 \leq \cdots \leq X_n$, which can be achieved by appropriately rearranging the samples. Define $r_i = \#\{j \mid Y_j \leq Y_i\}$ and $\ell_i = \#\{j \mid Y_j \geq Y_i\}$. Chatterjee (2021) proposed the following estimator:

$$\xi_n(X, Y) = 1 - \frac{n \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \sum_{i=1}^{n} \ell_i (n - \ell_i)},$$

and showed that, as n goes to infinity, it converges almost surely to the following measure of dependence first introduced by Dette et al. (2013):

$$\xi(X, Y) = \frac{\int_{-\infty}^{\infty} \operatorname{Var}\left[\mathrm{E}[\mathbf{1}_{\{Y \geq t\}} \mid X]\right] f_Y(t)\, dt}{\int_{-\infty}^{\infty} \operatorname{Var}\left[\mathbf{1}_{\{Y \geq t\}}\right] f_Y(t)\, dt}, \tag{1}$$

where f_Y is the PDF of Y and, for a proposition P, $\mathbf{1}_{\{P\}}$ is 1 if P is true and 0 if P is false.

We call $\xi(X, Y)$ the (asymmetric) *Chatterjee correlation coefficient*. We remark that some past works have used this nomenclature for the estimator ξ_n rather than for the measure of dependence ξ (Lin and Han, 2022b). However, our usage is consistent with more recent works (Bücher and Dette, 2024; Ansari and Fuchs, 2025; Ansari et al., 2025).

This coefficient is always bounded in $[0, 1]$ and, specifically, $\xi(X, Y) = 0$ if and only if X and Y are independent, and $\xi(X, Y) = 1$ if and only if Y is a measurable function of X (Chatterjee, 2021, Theorem 1.1). Note that the coefficient is not symmetric in general.

Chatterjee (2021) also proposed the following symmetric version of the coefficient:

$$\xi'(X,Y) = \max\{\xi(X,Y), \xi(Y,X)\}.$$

For this symmetric version, $\xi'(X,Y) = 1$ if and only if Y is a measurable function of X or X is a measurable function of Y.

A desirable property for a correlation coefficient, which is a similarity measure, is *transitivity*. Informally speaking, transitivity says that if variable X is correlated with variable Y, and variable Y is correlated with variable Z, then variable X should be (reasonably) correlated with variable Z. Transitivity of correlation coefficient makes the similarity relationship among variables both interpretable and coherent.

The *triangle inequality* can be seen as the distance-analog of transitivity. It says that for any three points (or variables) X, Y, Z, the distance from X to Z cannot exceed the sum of the distances from X to Y and from Y to Z. The triangle inequality is a highly desirable property of distance functions.

For instance, many clustering algorithms on metrics are able to provide guarantees on the quality of the returned clustering, thanks to the triangle inequality. Clustering is extremely useful in Machine Learning — in particular, it is used for feature selection tasks, as well as for grouping and/or nearest neighbor tasks. The Chatterjee correlation coefficients can naturally be transformed into the distance $d(X,Y) = 1 - \xi(X,Y)$ and $d'(X,Y) = 1 - \xi'(X,Y)$. This type of transformation has been extensively used in the literature, e.g., for the Jaccard coefficient/distance and the cosine similarity; see (Charikar, 2002). More recently, researchers have been studying the performance of clustering algorithms using distances based on Chatterjee's correlation coefficient (Fuchs and Wang, 2024).

In this paper, we study the metricity of the (directed) distance function obtained from the (asymmetric) Chatterjee coefficient, that is, we ask whether that distance satisfies the (directed) triangle inequality or, in other words, if the distance is a (quasi-)metric.

Our Contributions. Our main result is that the distance function implied by the symmetric Chatterjee coefficient, $d'(X,Y)$ is not a metric. Indeed we show that not only d' can violate the triangle inequality, but also that the violation can be arbitrarily large.

A more general way to obtain a distance function from the coefficient of correlation is to consider $d'_c(X,Y) = c - \xi'(X,Y)$, for some constant $c \geq 1$. We show that d'_c satisfies the triangle inequality if and only if $c \geq 2$.

A related question is whether the asymmetric distance function, $d(X,Y)$, is a quasi-metric. And, for the symmetric case, one can ask whether aggregators other than max make the resulting distance a metric. For example, $\min\{\xi(X,Y), \xi(Y,X)\}$ or $\frac{\xi(X,Y)+\xi(Y,X)}{2}$ are also symmetric coefficients which induce the symmetric distances $d'_{\min}$ and d'_{avg} in the natural way. We show that the triangle inequality can be violated, by a multiplicative constant, in each of these cases. Thus, we conclude that d is not a quasi-metric and $d'_{\min}$ and d'_{avg} are not metrics.

Experimentally, we find that all the distances d, d', $d'_{\min}$, and d'_{avg} violate the triangle inequality also in several real-world datasets.

Related Work. Chatterjee's coefficient and its estimator have attracted much attention. They have been used for tests of independence (Shi et al., 2021), for clustering tasks (Fuchs and Wang, 2024), for causal inference methods (Chatterjee and Vidyasagar, 2024), and they have been studied in combination with other correlation coefficients such as Spearman's (Zhang, 2024). The correlation coefficient has also been generalized using alternative construction principles (Ansari et al., 2025; Gao and Li, 2024) and adapted to test conditional dependence (Azadkia and Chatterjee, 2021). Properties of the correlation coefficient have also been investigated. Specifically, it has been shown that the coefficient is not continuous with respect to weak convergence (Bücher and Dette, 2024; Ansari and Fuchs, 2025); however, it respects a version of the data processing inequality (Chatterjee and Vidyasagar, 2024). Other estimators have also been proposed, with better theoretical convergence guarantees (Lin and Han, 2022b; Xia et al., 2024). Despite the great deal of attention, to the best of our knowledge, the metricity of the coefficient has not been studied before (Chatterjee, Personal communication, 2024).

The triangle inequality is not a necessary property for a distance measure but, nevertheless, it is desirable and has been extensively studied. In cases where the triangle inequality is violated, researchers have tried to identify alternatives which satisfy the triangle inequality, particularly for well-studied correlations such as cosine similarity (Schubert, 2021)

(van Dongen and Enright, 2012), as well as Pearson and Spearman correlations (van Dongen and Enright, 2012; Chen et al., 2023). Our work highlights how proving a violation of the triangle inequality for the Chatterjee correlation is non-trivial. Moreover, it shows that the straightforward approaches used to enforce the triangle inequality for cosine similarity, Pearson, and Spearman correlations (e.g., taking the square root) do not work for the Chatterjee correlation.

Notation. For random variables X, Y, we use the standard notation $F_{X,Y}(x, y) = \Pr[X \leq x, Y \leq y]$ for the joint cumulative distribution function (CDF) and $f_{X,Y}(x, y)$ for the joint probability density function (PDF). Recall that $f_{X,Y}$ is the derivative of $F_{X,Y}$ in the distributional sense. We use the analogous notation f_X, F_X for a single random variable. We use $\delta(\cdot)$ to denote the Dirac delta function. We use $X \perp Y$ to denote X and Y are independent. For $a > 0$, let $U(0, a)$ denote the uniform distribution on $[0, a]$. We use $X \sim D$ to denote that the random variable X is distributed according to D.

Organization. We prove our main result and its implications in Section 2. In Section 3 we prove our result on the asymmetric distance function d and in Section 4 we show that other symmetric functions, including d'_{avg} and $d'_{\min}$, also do not satisfy the triangle inequality. In Section 5 we report our experimental findings. We conclude with some final remarks in Section 6.

2 Symmetric Case

In this section, we show that the distance function $d'(X, Y)$ can violate the triangle inequality arbitrarily. We do so by exhibiting an explicit construction of three random variables. Specifically, we will prove the following:

Lemma 1. *Choose any $\epsilon \in (0, \frac{1}{2})$ and let $\alpha, \beta \sim U(0, \frac{1}{2})$, $\alpha \perp \beta$. Define:*

(i) $A = \mathbf{1}_{\{\alpha \geq \epsilon\}}$,

(ii) $B = \beta + \mathbf{1}_{\{\alpha \geq \epsilon\}}$,

(iii) $C = \alpha$.

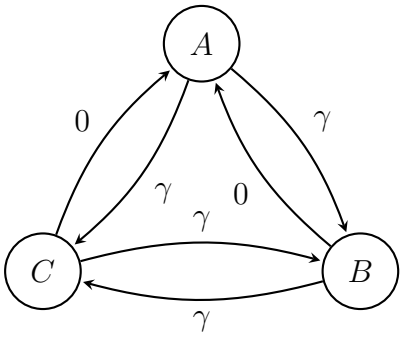


Figure 1: Construction of Lemma 1, $\gamma = 1-4\epsilon(1-2\epsilon) > 1-4\epsilon$. The weight of the directed arc (X, Y) is $d(X, Y)$. Note that $d'(X, Y) = \min\{d(X, Y), d(Y, X)\}$. Also note that the directed triangle inequality is satisfied in this construction.

Then, $\xi(C, A) = \xi(B, A) = 1$ *and* $\xi(A, C) = \xi(A, B) = \xi(B, C) = \xi(C, B) = 4\epsilon(1 - 2\epsilon) < 4\epsilon$.

The intuition behind the construction is that A is both a function of C and a function of B, therefore it will be at distance 0 from both of them using d'. However, B and C are almost independent, indeed, they are independent under the conditioning that $\alpha \geq \epsilon$, and thus, their distance will be close to 1. See Figure 1 for a visualization of the construction.

Before proving the lemma formally, we show its implications. We start by showing that $d'(X, Y)$ neither satisfies the triangle inequality nor any approximation of it.

Theorem 2. *For any* $\rho \in (0, 1)$*, there exists three random variables* A, B, C *such that* $d'(B, A) + d'(A, C) < \rho \cdot d'(B, C)$ *and* $d'(B, A) + d'(A, C) < d'(B, C) + \rho - 1$.

Proof. Consider the construction of Lemma 1 with $\epsilon = \frac{\rho}{4}$. Then, $d'(B, A) + d'(A, C) = 0$ and $d'(B, C) > 1 - 4\epsilon = 1 - \rho > 0$. The statement follows. □

Note that, not only $d'(X, Y)$ does not respect the triangle inequality, but it might also hold $d'(X, Y) = 0$ for two different random variables X, Y. Actually, the construction of Lemma 1 shows that the relation of being at distance 0 is not transitive. Note that the construction of Lemma 1 also shows that other transformations used for Spearman and Pearson correlation to obtain a metric (van Dongen and Enright, 2012), do not work for Chatterjee correlation, specifically, $\sqrt{1 - (\xi'(X, Y))^t}$, with $t \geq 1$, can also violate triangle inequality arbitrarily.

We can also obtain the minimum c such that $c - \xi'(X, Y)$ indeed respects the triangle inequality.

Theorem 3. *Let $d'_c(X, Y) = c - \xi'(X, Y)$, for $c \geq 1$. Then, d'_c satisfies the triangle inequality if and only if $c \geq 2$.*

Proof. If $c < 2$, consider the construction of Lemma 1 with $\epsilon = \frac{1}{2} - \frac{c}{4}$, then, $d'_c(B, A) + d'_c(A, C) = 2(c-1)$, while $d'_c(B, C) > c - 4\epsilon = 2(c-1)$. Instead, if $c \geq 2$, for any X, Y, Z, $d'_c(Y, X) + d'_c(X, Z) \geq 2(c-1) \geq c \geq d'_c(Y, Z)$. □

Now, we proceed to proving Lemma 1. The approach is to use the definition of the three random variables to explicitly compute the six Chatterjee coefficients. We will make use of the following simplification of the coefficient $\xi(X, Y)$ that holds when Y is continuous (i.e., for any $y \in \mathbb{R}$, $\Pr[Y = y] = 0$). This characterization was first given in (Dette et al., 2013, Theorem 2) and was proven to be equivalent to Equation (1) by Dalitz et al. (2024).

Theorem 4 (Theorem 1 of (Dalitz et al., 2024))**.** *For continuous random variable Y, it holds:*

$$\xi(X, Y) = 6 \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \Pr[Y \geq y \mid X = x]^2 \cdot f_X(x)\, dx \right) \cdot f_Y(y)\, dy - 2.$$

We now proceed to prove our Lemma.

Proof of Lemma 1. By construction A is a function of C, and hence by (Chatterjee, 2021, Theorem 1.1), we have that $\xi(C, A) = 1$. Similarly, A is a measurable function of B, indeed $A = \mathbf{1}_{\{B \geq 1\}}$ and hence $\xi(B, A) = 1$. Note that $f_C(c) = 2 \cdot \mathbf{1}_{\{0 \leq c \leq 1/2\}}$ and:

$$F_B(t) = \begin{cases} 4\epsilon t & \text{if } 0 \leq t \leq \frac{1}{2} \\ 2\epsilon & \text{if } \frac{1}{2} < t \leq 1 \\ 2\epsilon + (2 - 4\epsilon)(t-1) & \text{if } 1 < t \leq \frac{3}{2}, \end{cases} \quad \text{and} \quad f_B(t) = \begin{cases} 4\epsilon & \text{if } 0 \leq t \leq \frac{1}{2} \\ 2 - 4\epsilon & \text{if } 1 < t \leq \frac{3}{2} \\ 0 & \text{otherwise.} \end{cases}$$

We now compute $\xi(A, C)$ and $\xi(A, B)$. Since A is distributed as a Bernoulli distribution of parameter $1 - 2\epsilon$, by using Theorem 4 we have:

$$\xi(A, C) = 12 \int_0^{1/2} \left(2\epsilon \cdot \Pr[C \geq t \mid A = 0]^2 + (1 - 2\epsilon) \cdot \Pr[C \geq t \mid A = 1]^2 \right) dt - 2$$

$$= 12\left(\int_0^{\epsilon} 2\epsilon(\epsilon - t)^2\,dt + (1-2\epsilon)\left(\epsilon + \int_{\epsilon}^{1/2}\left(\frac{1-2t}{1-2\epsilon}\right)^2 dt\right)\right) - 2$$
$$= 12\left(\frac{2\epsilon^2}{3} + (1-2\epsilon)\left(\epsilon + \frac{1-2\epsilon}{6}\right)\right) - 2 = 4\epsilon(1-2\epsilon).$$

We can similarly compute $\xi(A,B)$:

$$\begin{aligned}\xi(A,B) &= \int_{-\infty}^{\infty}\left(2\epsilon\cdot\Pr[B\geq t\mid A=0]^2 + (1-2\epsilon)\cdot\Pr[B\geq t\mid A=1]^2\right) f_B(t)\,dt - 2\\ &= 24\epsilon\int_0^{1/2}\left(2\epsilon\cdot\Pr[\beta\geq t]^2 + (1-2\epsilon)\right)\,dt + 12(1-2\epsilon)^2\int_1^{3/2}\Pr[\beta\geq t-1]^2\,dt - 2\\ &= \frac{24\epsilon(1-2\epsilon)}{2} + \left(48\epsilon^2 + 12(1-2\epsilon)^2\right)\cdot\int_0^{1/2}(1-2t)^2\,dt - 2\\ &= 12\epsilon(1-2\epsilon) + 8\epsilon^2 + 2(1-2\epsilon)^2 - 2 = 4\epsilon(1-2\epsilon).\end{aligned}$$

Next we compute $\xi(B,C)$ and $\xi(C,B)$. We first derive the joint PDF $f_{C,B}(c,b)$. Specifically, for $b\in[0,\frac{3}{2}], c\in[0,\frac{1}{2}]$ it holds,

$$F_{C,B}(c,b) = \begin{cases} 4cb & \text{if } b\leq\frac{1}{2}, c<\epsilon\\ 4\epsilon b & \text{if } b\leq\frac{1}{2}, c\geq\epsilon\\ 2c & \text{if } b\geq 1, c<\epsilon\\ 4(b-1)(c-\epsilon)+2\epsilon & \text{if } b\geq 1, c\geq\epsilon\\ 0 & \text{otherwise.}\end{cases}$$

Thus,

$$f_{C,B}(c,b) = \begin{cases} 4 & \text{if } (b\leq\frac{1}{2},\ c<\epsilon) \text{ or } (b\geq 1,\ c\geq\epsilon)\\ 0 & \text{otherwise.}\end{cases}$$

Applying Theorem 4 we can calculate $\xi(B,C)$. We have that,

$$\begin{aligned}\xi(B,C) &= 12\int_0^{1/2}\int_{-\infty}^{\infty}\Pr\left[C\geq t\mid B=b\right]^2 f_B(b)\,db\,dt - 2\\ &= -2 + 48\epsilon\int_0^{1/2}\int_0^{1/2}\Pr\left[C\geq t\mid B=b\right]^2\,db\,dt + (24-48\epsilon)\int_0^{1/2}\int_1^{3/2}\Pr\left[C\geq t\mid B=b\right]^2\,db\,dt\\ &= -2 + 48\epsilon\int_0^{1/2}\int_0^{1/2}\left(\int_t^{\infty}\frac{f_{C,B}(c,b)}{f_B(b)}\,dc\right)^2 db\,dt + (24-48\epsilon)\int_0^{1/2}\int_1^{3/2}\left(\int_t^{\infty}\frac{f_{C,B}(c,b)}{f_B(b)}\,dc\right)^2 db\,dt\\ &= -2 + 48\epsilon\int_0^{\epsilon}\int_0^{1/2}\left(\int_t^{\epsilon}\frac{1}{\epsilon}\,dc\right)^2 db\,dt + (24-48\epsilon)\int_0^{1/2}\int_1^{3/2}\left(\int_{\max\{t,\epsilon\}}^{1/2}\frac{2}{1-2\epsilon}\,dc\right)^2 db\,dt\end{aligned}$$

$$
\begin{aligned}
&= -2 + \frac{48}{\epsilon}\int_0^{\epsilon}\int_0^{1/2}(\epsilon - t)^2\,db\,dt + \frac{24}{1-2\epsilon}\int_0^{1/2}\int_1^{3/2}(1 - 2\cdot\max\{t,\epsilon\})^2\,db\,dt \\
&= -2 + 8\epsilon^2 + \frac{12}{1-2\epsilon}\int_0^{1/2}(1 - 2\cdot\max\{t,\epsilon\})^2\,dt \\
&= -2 + 8\epsilon^2 + \frac{12}{1-2\epsilon}\left(1/6 - 2\epsilon^2 + 8/3\epsilon^3\right) \\
&= 4\epsilon(1-2\epsilon).
\end{aligned}
$$

Moreover, with a similar calculation we get that $\xi(C,B) = \xi(B,C)$. The proof is thus complete. □

3 Asymmetric Case

In this section, we present a construction that is weaker in terms of violation, but shows that the directed triangle inequality is violated for the distance function implied by the asymmetric Chatterjee coefficient, thus showing it is not even a quasi-metric. Interestingly, the construction in Lemma 1 satisfies the directed triangle inequality; see Figure 1. Instead, we will make use of a different construction that we analyze later.

Lemma 5. *Choose any $a \in (0, \frac{1}{2})$, and let $b = 1 - a$. Let $S, T \sim U(0,1)$, $S \perp T$. Define*

(i) $A = S$,

(ii) $B = \begin{cases} S & \text{if } S < a \\ a + (1-a)T & \text{if } S \geq a, \end{cases}$

(iii) $C = \begin{cases} b \cdot T & \text{if } S < b \\ S & \text{if } S \geq b. \end{cases}$

Then, $\xi(A,B) = \xi(B,A) = \xi(A,C) = \xi(C,A) = a(2-a)$ and $\xi(B,C) = \xi(C,B) = \frac{1-5a+8a^2-2a^3}{1-a}$.

Using this, we show the asymmetric distance d does not respect the directed triangle inequality.

Theorem 6. *There exists random variables A, B, C such that $d(B,A) + d(A,C) \leq 0.92 \cdot d(B,C)$, and $d(B,A) + d(A,C) < d(B,C) - 0.05$.*

Proof. Let A, B, C be the variables of Lemma 5, for some $a \in (0, 1/2)$. Let $f(a) = \frac{(1-a)^3}{a(a^2-4a+2)}$. Then,

$$\frac{d(B,A) + d(A,C)}{d(B,C)} = \frac{2(1-a(2-a))}{1 - \frac{1-5a+8a^2-2a^3}{1-a}} = f(a).$$

We have that $f'(a) = \frac{(1-a)^2(a^2+4a-2)}{a^2(a^2-4a+2)^2}$ and so f has a minimum at $a = \sqrt{6} - 2 \approx 0.45$. Moreover, $f(\sqrt{6}-2) \approx 0.919 \leq 0.92$. Similarly, let $g(a) = d(B,A) + d(A,C) - d(B,C) = \frac{2(1-2a)(a^2-3a+1)}{1-a}$. One can check that $g(\sqrt{6}-2) \approx -0.053 < -0.05$ (although $\sqrt{6}-2$ is not a minimum of g). The proof is thus complete. □

We now proceed to prove Lemma 5. As for the proof of Lemma 1, we explicitly compute all six Chatterjee coefficients using the definitions of the random variables.

Proof of Lemma 5. Since, for $t \in [0,1]$, $\Pr[B \leq t] = \Pr[C \leq t] = t$, we have $f_A(t) = f_B(t) = f_C(t) = \mathbf{1}_{\{0 \leq t \leq 1\}}$. We first compute all the joint PDFs, and then finish the computation of the Chatterjee coefficients. For $0 \leq u \leq 1$, $0 \leq v \leq 1$, we have,

$$F_{A,B}(u,v) = \begin{cases} \min\{u,v\} & \text{if } u < a \text{ or } v < a \\ a + \frac{(u-a)(v-a)}{1-a} & \text{if } u \geq a, v \geq a. \end{cases}$$

Thus, $f_{A,B}(u,v) = 0$ if $v < 0$ or $v > 1$ or $u < 0$ or $u > 1$, otherwise, for $0 \leq u, v \leq 1$, we have,

$$f_{A,B}(u,v) = \begin{cases} \delta(u-v) & \text{if } u < a \text{ or } v < a \\ \frac{1}{1-a} & \text{if } u \geq a, v \geq a. \end{cases}$$

For $0 \leq u, v \leq 1$, we have,

$$F_{A,C}(u,v) = \begin{cases} \frac{uv}{b} & \text{if } u < b, v < b \\ u & \text{if } u < b, v \geq b \\ v & \text{if } u \geq b, v < b \\ \min\{u,v\} & \text{if } u \geq b, v \geq b. \end{cases}$$

Thus, $f_{A,C}(u,v) = 0$ if $v < 0$ or $v > 1$ or $u < 0$ or $u > 1$, otherwise, for $0 \leq u, v \leq 1$, we have,

$$f_{A,C}(u,v) = \begin{cases} \frac{1}{b} & \text{if } u < b, v < b \\ \delta(u-v) & \text{if } u \geq b, v \geq b. \end{cases}$$

For $0 \le u, v \le 1$, by using that $b = 1 - a$, we have,

$$F_{B,C}(u,v) = \begin{cases} \frac{uv}{1-a} & \text{if } u < a, v < b \\ u & \text{if } u < a, v \ge b \\ \frac{av}{1-a} + \frac{1-2a}{1-a} \cdot \min\{v, u-a\} & \text{if } u \ge a, v < b \\ a + \frac{(u-a)(v-a)}{1-a} & \text{if } u \ge a, v \ge b. \end{cases}$$

Thus, by using that $b = 1 - a$, we have $f_{B,C}(u,v) = 0$ if $v < 0$ or $v > 1$ or $u < 0$ or $u > 1$, otherwise, for $0 \le u, v \le 1$, we have,

$$f_{B,C}(u,v) = \begin{cases} \frac{1}{1-a} & \text{if } u < a, v < b \\ \frac{1}{1-a} & \text{if } u \ge a, v \ge b \\ \frac{1-2a}{1-a} \cdot \delta(u-a-v) & \text{if } u \ge a, v < b. \end{cases}$$

We are now ready to compute the Chatterjee coefficients. Recall that, for $\alpha < \beta$, $\int_\alpha^\beta \delta(x-\gamma)\,dx = \mathbf{1}_{\{\alpha \le \gamma \le \beta\}}$, and $\delta(x) = \delta(-x)$. Thus, by Theorem 4,

$$\begin{aligned} \xi(A,B) &= -2 + 6\int_0^1\int_0^1 \left(\int_t^\infty f_{A,B}(u,v)\,dv\right)^2 du\,dt \\ &= -2 + 6\left(\int_0^1\int_0^a \left(\int_t^1 \delta(u-v)\,dv\right)^2 du\,dt + \int_0^1\int_a^1 \left(\int_t^\infty f_{A,B}(u,v)\,dv\right)^2 du\,dt\right) \\ &= -2 + 6\left(\int_0^1\int_0^a \mathbf{1}_{\{t \le u \le 1\}}\,du\,dt + \int_0^a\int_a^1 \left(\int_t^\infty f_{A,B}(u,v)\,dv\right)^2 du\,dt \right. \\ &\qquad \left. + \int_a^1\int_a^1 \left(\int_t^1 \frac{1}{1-a}\,dv\right)^2 du\,dt\right) \\ &= -2 + 6\left(\frac{a^2}{2} + \frac{(1-a)^2}{3} + \int_0^a\int_a^1 \left(\int_t^a \delta(u-v)\,dv + \int_a^1 \frac{1}{1-a}\,dv\right)^2 du\,dt\right) \\ &= -2 + 6\left(\frac{3a^2 + 2(1-a)^2}{6} + \int_0^a\int_a^1 du\,dt\right) = a(2-a). \end{aligned}$$

By a similar calculation, we have that $\xi(B,A) = a(2-a) = \xi(A,B)$. Next, we proceed to compute $\xi(A,C)$.

$$\begin{aligned} \xi(A,C) &= -2 + 6\int_0^1\int_0^1 \left(\int_t^\infty f_{A,C}(u,v)\,dv\right)^2 du\,dt \\ &= -2 + 6\left(\int_0^b\int_0^b \left(\int_t^b \frac{1}{b}\,dv\right)^2 du\,dt + \int_0^1\int_b^1 \left(\int_{\max\{b,t\}}^1 \delta(u-v)\,dv\right)^2 du\,dt\right) \end{aligned}$$

$$= -2 + 6\left(\int_0^b \frac{(b-t)^2}{b}\,dt + \int_0^1 \int_{\max\{b,t\}}^1 du\,dt\right)$$
$$= 1 - b^2 = a(2-a),$$

where we used that $b = 1-a$. By a similar calculation, we have $\xi(C,A) = a(2-a) = \xi(A,C)$. Finally,

$$\xi(B,C) = -2 + 6\int_0^1 \int_0^1 \left(\int_t^\infty f_{B,C}(u,v)\,dv\right)^2 du\,dt$$
$$= -2 + 6\left(\int_0^b \int_0^a \left(\int_t^b \frac{1}{1-a}\,dv\right)^2 du\,dt + \int_0^b \int_a^1 \left(\int_t^\infty f_{B,C}(u,v)\,dv\right)^2 du\,dt \right.$$
$$\left. + \int_b^1 \int_a^1 \left(\int_t^1 \frac{1}{1-a}\,dv\right)^2 du\,dt\right)$$
$$= -2 + 6\left(\frac{a(1-a)}{3} + \frac{a^3}{3(1-a)} + \int_0^b \int_a^1 \left(\int_t^b \frac{1-2a}{1-a}\cdot\delta(u-a-v)\,dv + \int_b^1 \frac{1}{1-a}\,dv\right)^2 du\,dt\right)$$
$$= -2 + 6\left(\frac{a(1-a)^2+a^3}{3(1-a)} + \int_0^b \int_a^1 \left(\frac{(1-2a)\mathbf{1}_{\{t+a\le u\le 1\}} + a}{1-a}\right)^2 du\,dt\right)$$
$$= -2 + 6\left(\frac{a(1-a)^2+a^3}{3(1-a)} + \int_0^b \left(\int_a^{t+a} \frac{a^2}{(1-a)^2}\,du + \int_{t+a}^1 du\right)\,dt\right)$$
$$= -2 + 6\left(\frac{a(1-a)^2+a^3}{3(1-a)} + a - \frac{1}{2} + (1-a)^2\right)$$
$$= \frac{1-5a+8a^2-2a^3}{1-a}.$$

With a similar calculation, we also get $\xi(C,B) = \xi(B,C)$. □

4 Other Symmetrizations

While we have shown neither the symmetric distance d' nor the asymmetric distance d is a metric, a natural question is to ask if there are other ways to define a symmetric version based on d that might make it a metric.

A natural way to construct other symmetric versions of the coefficient is the following. Let $\Phi : [0,1]^2 \to \mathbb{R}_{\ge 0}$ be a symmetric function. Define the generalized symmetric distance as $d'_\Phi(X,Y) = \Phi(1-\xi(X,Y), 1-\xi(Y,X)) = \Phi(d(X,Y), d(Y,X))$.

Theorem 7. *Consider a symmetric* $\Phi : [0,1]^2 \to \mathbb{R}_{\geq 0}$ *and suppose that there exists* $a \in (0, \frac{1}{2})$ *such that for* $v_1 = (1-a)^2$ *and* $v_2 = \frac{2a(a^2-4a+2)}{1-a}$ *it holds that* $2 \cdot \Phi(v_1, v_1) < \Phi(v_2, v_2)$. *Then, there exists random variables* A, B, C *such that* $d'_\Phi(B, A) + d'_\Phi(A, C) < d'_\Phi(B, C)$.

Proof. Consider the random variables A, B, C of Lemma 5 with the parameter a of the statement. Then, $d'_\Phi(B, A) = d'_\Phi(A, C) = \Phi(v_1, v_1)$ and $d'_\Phi(B, C) = \Phi(v_2, v_2)$. Thus, $d'_\Phi(B, A) + d'_\Phi(A, C) = 2 \cdot \Phi(v_1, v_1) < \Phi(v_2, v_2) = d'_\Phi(B, C)$. □

In particular, we consider the following two natural ways to define a symmetric notion of distance: $d'_{\min}(X, Y) = 1 - \min\{\xi(X, Y), \xi(Y, X)\}$ and $d'_{\text{avg}}(X, Y) = 1 - \frac{\xi(X,Y)+\xi(Y,X)}{2}$. It is easy to see that Theorem 7 applies to these two distances for, e.g., $a = \sqrt{6} - 2$.

Differently from using the maximum function, the relation of being at distance 0 is transitive for these two distances. Indeed, by (Chatterjee, 2021, Theorem 1.1), $d'_{\min}(X, Y) = 0$ and $d'_{\text{avg}}(X, Y) = 0$ if and only if $X = f(Y)$ and $Y = g(X)$ for two measurable functions f, g (see also (Fuchs and Wang, 2024, Theorem 2.7)). Therefore, if $d'_{\min}(X, Y) = d'_{\min}(Y, Z) = 0$ then $d'_{\min}(X, Z) = 0$ because the composition of measurable functions in the same σ-algebra is also measurable. Then, one might hope to obtain a metric by restricting to the equivalence classes of the relation of being at distance 0, but, unfortunately, the same construction of Lemma 5 shows that these distance functions on the equivalence classes also do not satisfy the triangle inequality. Indeed, the random variables A, B, C of Lemma 5 belong to different equivalence classes, being at distance larger than 0 from each other. We mention that Siburg and Stoimenov (2010) also studied symmetric measures of dependence that evaluate to 1 if and only if X is a function of Y and Y is a function of X.

Our constructions also applies to generalizations of the distance d'. Formally, we have:

Theorem 8. *Consider a symmetric* $\Phi : [0,1]^2 \to \mathbb{R}_{\geq 0}$ *and suppose that* $\Phi(0,0) = \Phi(u,0) = \Phi(0,v) = 0$ *and* $\Phi(u,v) > 0$ *for all* $u, v \in (0,1]^2$. *Then, there exists random variables* A, B, C *such that* $d'_\Phi(B, A) + d'_\Phi(A, C) < d'_\Phi(B, C)$.

Proof. Consider the random variables A, B, C of Lemma 1 with $\epsilon = \frac{1}{4}$. Then, $d'_\Phi(C, A) = \Phi(0, \frac{1}{2}) = 0$ and $d'_\Phi(A, B) = \Phi(\frac{1}{2}, 0) = 0$. Instead, $d'_\Phi(B, C) = \Phi(\frac{1}{2}, \frac{1}{2}) > 0$. □

We highlight that Theorem 8 applies to the class of distances considered by Fuchs and Wang (2024) for clustering tasks in their Corollary 2.3.(A), showing that none of the

distances in such family is a metric.

5 Experiments

In this section, we aim to answer the following question: while the distance measures implied by the Chatterjee coefficient are not metrics in theory, is the triangle inequality violated in practice? To answer this question, we consider many real-world datasets, consisting of several random variables, and we compute two quantities: the first captures the magnitude of the violations, and the second captures how many violations arise in the datasets. Specifically, first, for each dataset, we compute the worst possible additive violation of the triangle inequality (if any violation exists). The second quantity consists in computing, for each $x \in \{1, \dots, 100\}$, in how many datasets the $x\%$ of undirected and directed triples violate the triangle inequality.

5.1 Experimental Setup

To compute the Chatterjee correlation coefficient, $\xi(X, Y)$, in practice, we use the estimator $\xi_n(X, Y)$ proposed by Chatterjee (2021), which can be computed efficiently and using only a finite number of samples from the joint distribution of (X, Y).

Hyperparameters. We set $n = 5 \cdot 10^4$, which is of the same order of magnitude as the parameter used by Chatterjee (2021). We repeat each computation of ξ_n five times, and report confidence intervals with respect to these five runs.

Datasets. We use 80 datasets from the UCI Machine Learning Repository (Kelly et al., 2024). Each of these datasets is meant to be used for classification, regression, or clustering tasks and consists of features and instances. For our purposes, each feature is a random variable, and to obtain a draw from the joint distribution, we sample a uniform at random instance and return the corresponding feature values. We use only continuous features without missing values. All datasets are publicly available in the UCI Python package, which is released under the MIT license.[1] More precisely, we used all the datasets available in the

[1] `https://github.com/uci-ml-repo/ucimlrepo`

Python package with a number of features between 3 and 100, for computational reasons. These datasets span from 1993 to 2024 and contain some well-known machine learning datasets such as Abalone (Nash et al., 1994) and Breast Cancer Wisconsin (Wolberg et al., 1993), just to name a few.

The experiments were done on a standard desktop computer.

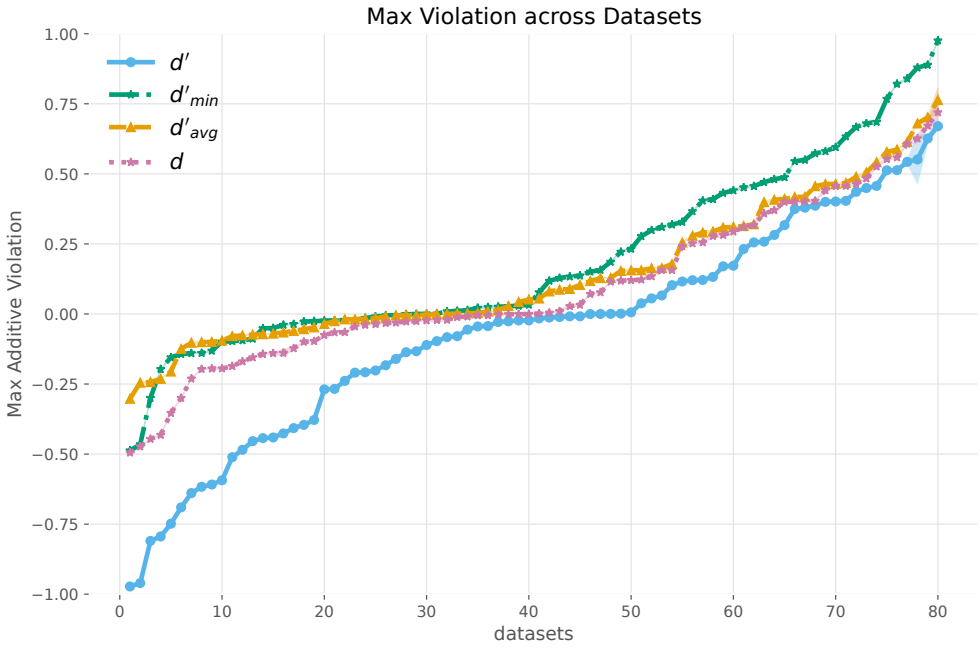


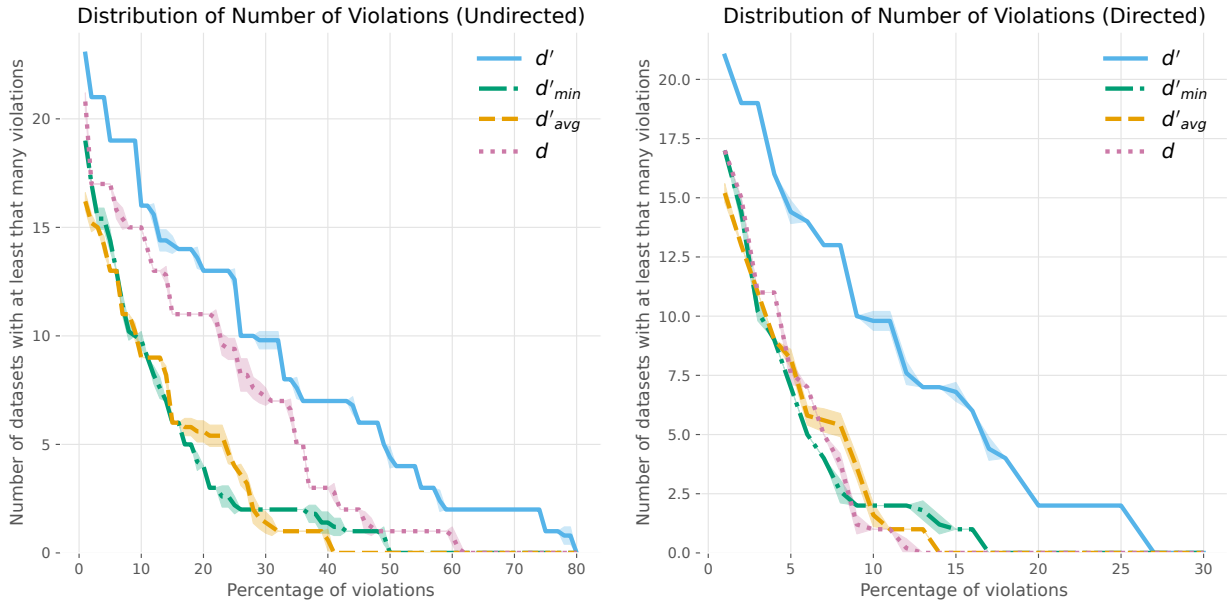
Figure 2: For each dataset, we report the worst possible additive violation of the triangle inequality. The line of each distance is sorted in increasing order, independently of the others. The standard deviation with respect to the random seed is negligible.

5.2 Maximum Violation

For a dataset consisting of random variables $X_1, \ldots, X_N$, we compute the worst possible additive violation of the triangle inequality, i.e., for a distance $D \in \{d, d', d'_{\text{avg}}, d'_{\text{min}}\}$,

$$\min_{\substack{i,j,k \in [N], \\ i \neq j, i \neq k, j \neq k}} D(X_i, X_j) + D(X_j, X_k) - D(X_i, X_k).$$

Note that if this value is positive, the dataset satisfies the triangle inequality. We report in Figure 2 the results for all the datasets. Note that distance d', that uses max to symmetrize the coefficient, reaches a violation of almost -1, the worst possible, for two datasets.



(a) Distribution for *undirected* triples.

(b) Distribution for *directed* triples.

Figure 3: For each possible percentage of violated triples, we report the number of datasets violating at least that fraction of triples. The shaded area corresponds to one standard deviation with respect to the random seed. The maximum number of datasets is 38 in these plots.

Moreover, most datasets (46 out of 80) violate triangle inequality in at least one triple using d'. The three other distances are generally more robust than d' to violations of triangle inequality, but they can still reach a violation as low as -0.5. Overall, triangle inequality can be badly violated in real-world datasets.

5.3 Distribution of the Violations

In this section, we aim to understand how many triples of random variables violate the triangle inequality in the datasets. We consider both undirected/unordered and directed/ordered triples. More precisely, we say that an undirected triple of random variables violates the triangle inequality if there exists an ordering of the random variables for which they violate the triangle inequality. More formally, $\{X_1, X_2, X_3\}$ violates the triangle inequality if the

following quantity is negative,

$$\min_{\pi \in \mathbf{S}_3} D(X_{\pi(1)}, X_{\pi(2)}) + D(X_{\pi(2)}, X_{\pi(3)}) - D(X_{\pi(1)}, X_{\pi(3)}),$$

where $\mathbf{S}_3$ is the set of permutations on three elements, and $D \in \{d, d', d'_{\text{avg}}, d'_{\min}\}$. We instead say that a directed triple (X_1, X_2, X_3) violates the triangle inequality if $D(X_1, X_2) + D(X_2, X_3) - D(X_1, X_3)$ is negative.

Since the datasets have different number of features (i.e., random variables), we look at the fraction of triples violated in a dataset rather than the total number. More precisely, fixed a percentage of violated triples $x \in \{1, \ldots, 100\}$, we compute, for undirected triples, the number of datasets such that,

$$\frac{100}{\binom{N}{3}} \sum_{i<j<k \in [N]} \mathbf{1}_{\{\{X_i, X_j, X_k\} \text{ is violated}\}} \geq x,$$

where the considered dataset has random variables $\{X_1, \ldots, X_N\}$. For directed triples we perform the analogous check,

$$\frac{100}{N(N-1)(N-2)} \sum_{\substack{i,j,k \in [N], \\ i \neq j, i \neq k, j \neq k}} \mathbf{1}_{\{(X_i, X_j, X_k) \text{ is violated}\}} \geq x.$$

Several datasets in the 80 that we consider have only a few number of random variables. This would artificially boost the percentage of violated triples. Thus, in this part of the experiment, we restrict to the 38 datasets that have more than 7 features (the median value). We report the results in Figure 3.

We can see that, for undirected triples and using distance d', there are datasets for which more than 70% of the triples are violated. Again, the other distances are more robust but they can still reach more than 50% violated triples. The number of violated directed triples is substantially smaller, but still significant: with all four distances there are datasets with more than 10% violations.

6 Conclusion

We showed that the symmetric distance, d', implied by the Chatterjee correlation coefficient and by using the max function to symmetrize it, does not respect the triangle inequality,

not even approximately. We also showed that none of d, d'_{avg}, d'_{min} respects the triangle inequality. However, our second construction does not rule out the possibility that d, d'_{avg}, or d'_{min} satisfy an appproximate vesion of the triangle inequality. This is a natural direction to further investigate and we leave it as an open question. Our experiments suggest that the additive violation of Theorem 6 is not optimal. However, since the datasets consists of many instances, the random variables that we used in the experiments are discrete with a very large support, and are therefore analyzable only experimentally.

While we showed that none of d, d', d'_{avg}, d'_{min} satisfies the triangle inequality in general, it might still be that these distances are metrics when restricted to some classes of distributions. This is another interesting direction to explore.

Finally, one can think of other ways to obtain a symmetric version of the coefficient for which our constructions do not apply. For instance, Fuchs and Wang (2024) considered $1 - \xi(X, Y) \cdot \xi(Y, X)$ and $1 - \max\{0, \xi(X, Y) + \xi(Y, X) - 1\}$. We leave open the problem of classifying the metricity of these and other distances.

Data and code availability. The code to reproduce the experiments is available on GitHub at: `https://github.com/mirkogiacchini/metricityChatterjeeCoefficient`. The data is publicly available at: `https://github.com/uci-ml-repo/ucimlrepo`.

Acknowledgements. We thank Prof. Sourav Chatterjee, the anonymous reviewers, and the anonymous associate editor for useful comments.

Funding. Flavio Chierichetti was supported in part by BiCi – Bertinoro international Center for informatics, by the PRIN project 20229BCXNW, and by a Google Focused Research Award.

Competing interests. The authors report there are no competing interests to declare.

References

Ansari, J. and S. Fuchs (2025). On continuity of Chatterjee's rank correlation and related dependence measures. *arXiv 2503.11390*.

Ansari, J., P. B. Langthaler, S. Fuchs, and W. Trutschnig (2025). Quantifying and estimating dependence via sensitivity of conditional distributions. *Bernoulli*.

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

Bücher, A. and H. Dette (2024). On the lack of weak continuity of Chatterjee's correlation coefficient. *arXiv 2410.11418*.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on the Theory of Computing*, pp. 380–388.

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

Chatterjee, S. and M. Vidyasagar (2024). Estimating large causal polytrees from small samples. *arXiv 2209.07028*.

Chen, J., Y. K. Ng, L. Lin, X. Zhang, and S. Li (2023). On triangle inequalities of correlation-based distances for gene expression profiles. *BMC Bioinformatics 24*(1), 40.

Dalitz, C., J. Arning, and S. Goebbels (2024). A simple bias reduction for Chatterjee's correlation. *Journal of Statistical Theory and Practice 18*.

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.

Fuchs, S. and Y. Wang (2024). Hierarchical variable clustering based on the predictive strength between random vectors. *International Journal of Approximate Reasoning 170*.

Gao, M. and Q. Li (2024). A family of Chatterjee's correlation coefficients and their properties. *arXiv 2403.17670*.

Kelly, M., R. Longjohn, and K. Nottingham (2024). The UCI machine learning repository. https://archive.ics.uci.edu. Accessed: 10/09/2024.

Lin, Z. and F. Han (2022a). Limit theorems of Chatterjee's rank correlation. *arXiv 2204.08031*.

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

Nash, W., T. Sellers, S. Talbot, A. Cawthorn, and W. Ford (1994). Abalone. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55C7W.

Schubert, E. (2021). A triangle inequality for cosine similarity. In *Proceedings 14th International Conference on Similarity Search and Applications*, pp. 32–44.

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

Siburg, K. F. and P. A. Stoimenov (2010). A measure of mutual complete dependence. *Metrika 71*, 239–251.

van Dongen, S. and A. J. Enright (2012). Metric distances derived from cosine similarity and Pearson and Spearman correlations. *arXiv 1208.3145*.

Wolberg, W., O. Mangasarian, N. Street, and W. Street (1993). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.

Xia, L., R. Cao, J. Du, and X. Chen (2024). The improved correlation coefficient of Chatterjee. *Journal of Nonparametric Statistics 37*(2), 265–281.

Zhang, Q. (2024). On relationships between Chatterjee's and Spearman's correlation coefficients. *Communications in Statistics - Theory and Methods 54*(1), 259–279.