The coverage correlation coefficient: Going beyond functional dependence.

Xuzhi Yang, Mona Azadikia, Tengyao Wang

Department of Statistics, LSE

Geomtry of Chatterjee's correlation coefficient

Given random samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^{(X,Y)} \in \mathcal{P}(\mathbb{R}^{d_X+d_Y})$ with $d_X, d_Y \geq 1$. The goal is to construct a coefficient of correlation to measure the dependency between random vectors X and Y.

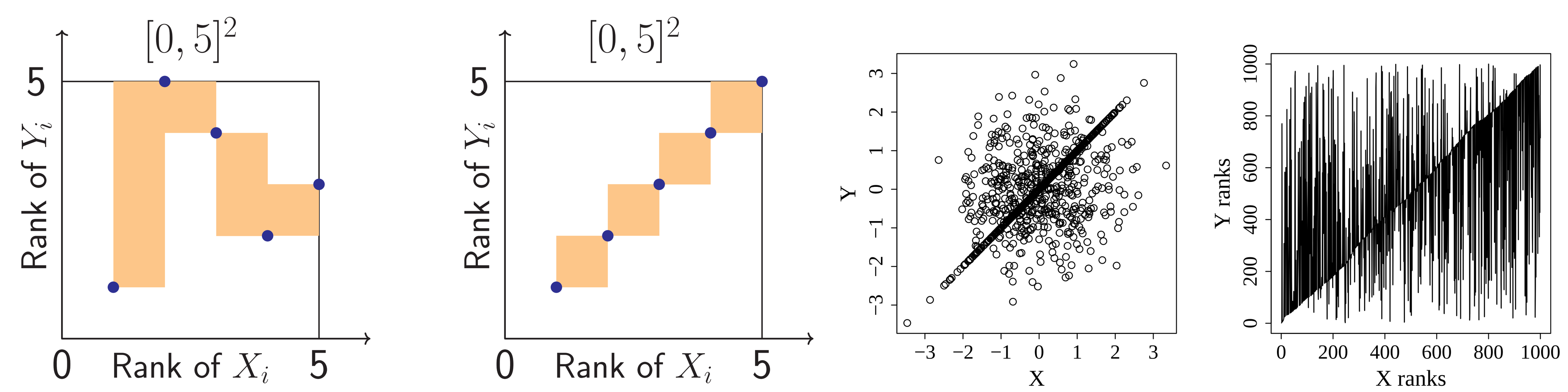
Chatterjee's correlation coefficient ($d_X = d_Y = 1$): Assume $X_1 < X_2 < \cdots < X_n$, let $r_i = \mathrm{Rank}(Y_i)$ for $i = 1, \ldots, n$. Let

$$\xi_n := 1 - \frac{\sum_{i=1}^{n} |r_{i+1} - r_i|}{(n^2-1)/3}. \tag{1}$$

It is shown that ξ_n

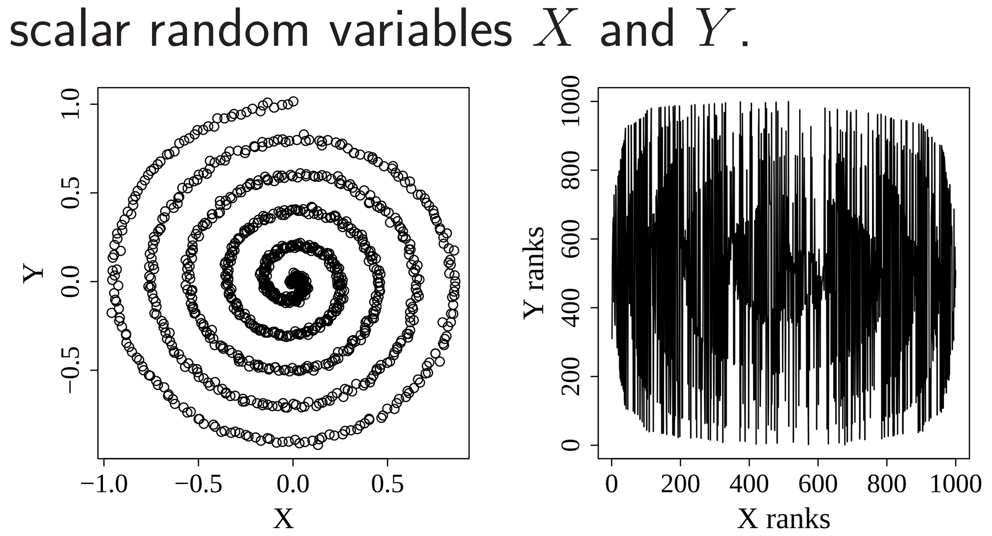
- consistently estimates a population quantity equals to 0 if and only if $X \perp\!\!\!\perp Y$, and equals to 1 if and only if $Y = f(X)$ for some measurable function f;
- allows a distribution-free null asymptotic theory;
- can be computed in time $O(n \log n)$.

Geometric intuition: let $n = 5$, we visualise the numerator of (1) as follows:



⇒ Independence data can induce larger covered area!

- However, it can cover to much area under non-functional correlation (see the figure below);
- It only computes correlation of scalar random variables X and Y.



Coverage correlation coefficient

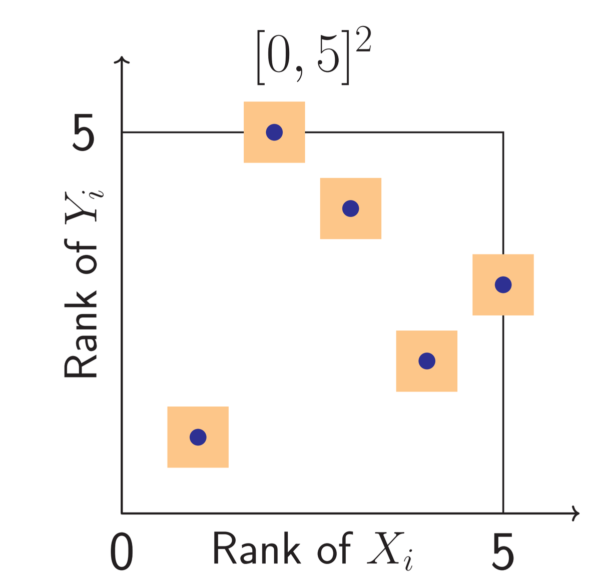
In this work we propose a new coefficient of correlation that

- can measure the dependence between random vectors;
- enjoys a distribution-free null asymptotic theory;
- consistently estimate a population quantity equals to 0 if and only if $P^{(X,Y)} = P^X \otimes P^Y$, equals to 1 if and only if $P^{(X,Y)} \perp P^X \otimes P^Y$;
- allows a $O(n \log n)$ algorithm under the case of univariate marginals.

Coverage correlation coefficient

1. Let the optimal transport-based rank of X_i and Y_i be $\widehat{R}^X(X_i)$ and $\widehat{R}^Y(Y_i)$, respectively. Define the joint rank $\widehat{R}_i = (\widehat{R}^X(X_i), \widehat{R}^Y(Y_i))$, for $i = 1, \ldots, n$.
2. For each $\widehat{R}_i$, we construct a small cube centred at $\widehat{R}_i$ with edge length $n^{-1/d}$, where $d = d_X + d_Y$.
3. Let V_n be the vacancy area/volume, define

$$\kappa_n^{X,Y} := \frac{V_n - e^{-1}}{1 - e^{-1}}$$



Consistency and asymptotic normality

Theorem . *Define $f : \mathbb{R} \to \mathbb{R}$ as $f(x) = (e^{-x} - e^{-1})/(1 - e^{-1})$. Then, when $d_X = d_Y = 1$ we have*

$$\kappa_n^{X,Y;\star} \xrightarrow{\text{p}} \kappa^{X,Y} := D_f(P^{(X,Y)} \,\|\, P^X \otimes P^Y) \quad \text{as } n \to \infty,$$

for $\star \in \{\mathrm{Reg}, \mathrm{Rand}\}$.

- $D_f(P^{(X,Y)} \,\|\, P^X \otimes P^Y) = 1$ iff. $P^{(X,Y)} \perp P^X \otimes P^Y$, where the functional dependence is a special case.

Theorem . *Let $P^X \in \mathcal{P}(\mathbb{R}^{d_X})$, $P^Y \in \mathcal{P}(\mathbb{R}^{d_Y})$. Then when X and Y are independent, we have*

$$n^{1/2} \kappa_n^{X,Y;\mathrm{Rand}} \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2), \quad \text{as } n \to \infty,$$

where $\sigma^2 = (e-1)^{-2} \sum_{i=2}^{\infty} \frac{1}{i!} \left(\frac{2}{i+1}\right)^d$.

- In practice, the explicit mean and variance for $\kappa_n^{X,Y;\mathrm{Rand}}$ is available by

$$\mathrm{E}(V_{n,\gamma}^{\mathrm{Rand}}) = (1 - 1/n)^n, \quad \mathrm{Var}(V_{n,\gamma}^{\mathrm{Rand}}) = \sum_{r=2}^{n} \binom{n}{r} \left(1 - \frac{2}{n}\right)^{n-r} \left(\left(\frac{2}{r+1}\right)^d n^{-r-1} - n^{-2r}\right).$$

Simulations