Exploring Cross Correlation Using Shuffles of M

Aishee Bhattacharya
Bipayan Banerjee
Rahul Konar
Srijeet Bhattacharjee

May 12, 2024

Abstract

Intuition says that two quantities being dependent on each other and two quantities being independent of each other are the polar opposite of each other. Whereas in one case, one quantity can be perfectly and confidently deduced from the other, in the other case, one quantity does not provide any significant information about the other. But, in reality, are the two cases so different?

On one hand there is a theorem due to Nelson which roughly states that independent random variables can be approximated arbitrarily by dependent random variables using ”Shuffles of M”.

On the other hand, The Chatterjee Correlation Coefficient aims to be able to detect all kinds of dependence and independence.

This project analyses how the Chaterjee Correlation coefficient behaves with respect to these shuffles of M, and try to see if dependence is really the other side of the coin or dependence and independence are more closer than we think.

Introduction

Basic Definitions

1. **Joint Distribution Function**

   A Joint Distribution Function H is a function over domain $\mathbb{R}^2$ such that the following properties hold -

   (a) H is **2-increasing** i.e. for $(x_1, x_2), (y_1, y_2) \in \mathbb{R}^2$ and $x_1 < y_1, x_2 < y_2$
   $\Delta H_{x,y} = H(x_2, y_2) - H(x_1, y_2) - H(x_2, y_1) + H(x_1, y_1) \geq 0$.

   (b) $\lim\limits_{x \to -\infty} H(x, y) = \lim\limits_{y \to -\infty} H(x, y) = 0$ and $\lim\limits_{x \to \infty} \lim\limits_{y \to \infty} H(x, y) = 1$

   (c) $(a_n, b_n) \downarrow (a, b) \implies H(a_n, b_n) \to H(a, b)$

   Probabilistically, $H(x, y) = P(X \leq x, Y \leq y)$, where X,Y are two random variables which are said to have the distribution H.

2. **Independent Random Variable**

   Two random variables X and Y are said to be independent if $H(x, y) = F(x)G(y)$ where H is the joint distribution function and F , G are the marginal distributions of X and Y respectively.

3. **Completely Dependent Random Variable**

   A random variable Y is said to be *completely dependent* on X if there exists a real function g such that $P[Y = g(X)] = 1$.

4. **Mutually Completely Dependent Random Variable**

   Two random variables X and Y are said to be *Mutually completely dependent* if there exists a bijection f such that $P[Y = f(x)] = 1$ i.e. X and Y are almost surely invertible functions of one another.

   Note that if $P[Y = X^2] = 1$ then Y is completely dependent on X but not mutually completely dependent since $h(x) = x^2$ is not a bijection.

5. **Support of a Random Variable**

   The support of a random variable X is the smallest closed set S such that $P(X \in S) = 1$ holds. We also have $x \in S$ iff for all $a < x < b$, $F(a) < F(b)$ i.e. for all $\epsilon > 0, P((x - \epsilon, x + \epsilon)) > 0$.

Nelsen's Theorem

Nelsen, in his book, **Introduction to Copulas** discusses a theorem, arguing that given independent random variables X and Y, following certain distributions, for arbitrary $\epsilon > 0$ one can construct random variables U and V, where

- $U \sim X$ and $V \sim Y$
- $V = f(U)$ where f is some real valued function.
- $$|F_{UV}(u, v) - F_{XY}(x, y)| < \epsilon$$

  where F_{UV} is the joint distribution function of U and V, and F_{XY} is the joint distribution function of X and Y.

Copula

A two-dimensional **Copula** is a function $C : \mathbf{I}^2 \to \mathbf{I}$ such that the follwing properties hold -

1. For every $u, v \in \mathbf{I}$

   $$C(u, 0) = 0 = C(0, v)$$

   and

   $$C(u, 1) = u \text{ and } C(1, v) = 1$$

2. For every u_1, u_2, v_1, v_2 in $\mathbf{I}$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

   $$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

Here is another formulation for Copula which may help in the proof of Nelsen's theorem later. If we denote $\Delta H_{x,y}$ by $V_H(B)$ where B is the set $[x_1, x_2] \times [y_1, y_2]$, then $C(u, v) = V_C([0, u] \times [0, v])$.

Now we present here an important lemma related to copulas.
Lemma : Let C be a copula then for all $(u_1, u_2), (v_1, v_2) \in \mathbf{I}^2$

$$|C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1| + |v_2 - v_1|$$

Sklar's Theorem

Let H be a joint distribution function with margins F and G. Then there exists a copula C such that for all $x, y \in \mathbb{R}$

$$H(x, y) = C(F(x), G(y))$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $RanF \times RanG$. Conversely, if C is a copula and F and G are distribution functions, then the function H defined above is a joint distribution function with margins F and G.

Now, from Sklar's theorem, we notice that any joint distribution function can be written as a copula of the form $C(F(x), G(y))$, where $F(x)$ and $G(y)$ are the distribution functions of the randomm variables X and Y. So in the statement of Nelsen's theorem the expression $|F_{UV}(u,v) - F_{XY}(x,y)|$ can be replaced by the expression $|C_{U,V}(u,v) - \Pi_{X,Y}(x,y)|$. Hence proving that $|C_{U,V}(u,v) - \Pi_{X,Y}(x,y)| < \epsilon$ is enough.

Two Important Copulas

1. **Fréchet-Hoeffding Upper Bound(M)-** We now give definition of this upper bound , but before that we present here a theorem here :

   Theorem Let C be a copula , then for every (u,v) in $\mathbf{I}^2$

   $$max(u+v-1,0) \leq C(u,v) \leq min(u,v).$$

   Here notice that $M(u,v) = min(u,v)$ and $W(u,v) = max(u+v-1,0)$ are themselves copulas.

   $M(u,v)$ is called the Fréchet-Hoeffding upper bound and $W(u,v)$ is called the Fréchet Hoeffding lower bound.

   Some interesting facts about Fréceht Hoeffding upper bound are -

   1. It dominates any other joint distribution function pointwise.

   2. It's support is a non-decreasing subest of $\mathbb{R}^2$.

   In particular, the joint distrn function of (X,Y) with X=Y and X any marginal distrn, gives this copula. Note that here X,Y are mutually dependent.

2. **Indpendence(Π) -**

   Let X and Y be continuous random variables. Then X and Y are independent iff $C_{XY} = \Pi$, where $\Pi(u,v) = uv$. Π is called the product copula.

Shuffles of M

The mass distribution for a shuffle of M is obtained by placing the mass of M on $\mathbf{I}^2$ and then $\mathbf{I}^2$ is vertically cut into a finite number of strips. Thereafter those strips are permuted/shuffled(even some of them can be flipped around their vertical axes of symmetry) and reassembled to form the square again. The resulting mass distribution will correspond to a copula called a shuffle of M.

Formally, a shuffle of M is determined by a positive integer n, a finite partition $J_i = J_1, J_2, \ldots, J_n$ of $\mathbf{I}$ into n closed subintervals, a permutation π on $S_n = 1, 2, \ldots, n$ and a function $\omega : S_n \to \{-1, 1\}$, where $\omega(i)$ is -1 or 1 according to whether or not the strip $J_i \times \mathbf{I}$ is flipped. We denote the shuffles/permutations of the strips by the vector of images $(\pi(1), \ldots, \pi(n)$.

Analytic Proof

Now we come to the analytic proof of Nelsen's theorem. The statement of the theorem is

For any $\epsilon > 0$, there exists a shuffle of M, which we denote by C_ϵ, such that

$$\sup_{u,v \in \mathbf{I}} |C_\epsilon(u,v) - \Pi(u,v)| < \epsilon$$

Now we come to the proof.

Proof By lemma 1 we know that for any copula C and for u, v, s, t in $\mathbf{I}$

$$|C(u,v) - C(s,t)| < |u-s| + |v-t|$$

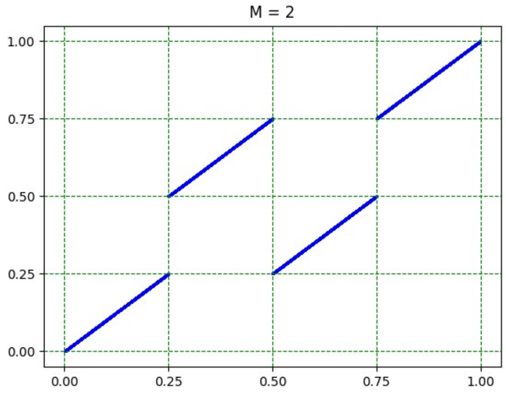
Now we can find an integer m such that $m \geq \frac{4}{\epsilon}$ i.e. $\dfrac{1}{m} \leq \dfrac{\epsilon}{4}$.

So whenever $|u-s| < \frac{1}{m}$ and $|v-t| < \frac{1}{m}$, $|C(u,v) - C(s,t)| < |u-s| + |v-t| < \frac{1}{m} + \frac{1}{m} = \frac{2}{m} \leq \frac{\epsilon}{2}$ i.e. we have got

$$|C(u,v) - C(s,t)| < \frac{\epsilon}{2}$$

.

We now choose m such that it determines C_ϵ , a shuffle of M in the following way : Let $n = m^2$, and let $\{J_i\}$ be the regular partition I_n of I into n subintervals of equal width i.e. of width $\frac{1}{n}$. Let π be the permutation of S_n given by $\pi(m(j-1)+k) = m(k-1)+j$ for $k, j = 1, 2, \dots, m$. What this permutation essentially does is redistributing the the probability mass of M so that there is mass $\frac{1}{n}$ in each of the n subsquares of $\mathbf{I}^2$. The figure below illustrates shuffle of M for $m = 2$.



We also notice that $V_{C_\epsilon}([0, \frac{p}{m}] \times [0, \frac{q}{m}]) = V_\Pi([0, \frac{p}{m}] \times [0, \frac{q}{m}]) = \frac{pq}{m^2} = \frac{pq}{n}$ for $p, q = 0, 1, \dots, m$.

Now let's take an arbitrary point (u, v) in $\mathbf{I}^2$. Then this point lies in any one of the n sub-squares of $\mathbf{I}^2$ i.e. there exists a pair of integers (p, q) with $p, q \in \{0, 1, \dots, m\}$ such that $|u - \frac{p}{m}| < \frac{1}{m}$ and $|v - \frac{q}{m}| < \frac{1}{m}$. Hence

$$|C_\epsilon(u,v) - \Pi(u,v)| \leq |C_\epsilon(u,v) - C_\epsilon(\frac{p}{m}, \frac{q}{m})| + |C_\epsilon(\frac{p}{m}, \frac{q}{m}) - \Pi(\frac{p}{m}, \frac{q}{m})| + |\Pi(\frac{p}{m}, \frac{q}{m}) - \Pi(u,v)| < \frac{\epsilon}{2} + 0 + \frac{\epsilon}{2} = \epsilon$$

Hence the proof is quite elegantly done!

Note - Note, there is nothing special about $\prod$ and any arbitrary copula can be generated through approximations by Shuffles of M. In fact, it can be shown that the set of Shuffles of M is dense in the space of all Copulas with sup norm.

Algorithm

Generation from Uniform Distribution

We denote $m^2; (m \in \mathbb{N})$ to be the number of partitions of $[0, 1]$. We generate $\mathbf{X} \in \mathbb{R}^n$, each coordinate from $U(0, 1)$ and Y vector by $Y_i = f_M(X_i)$, where f_M is the function which gives the M^{th} shuffle of M. We generate it using the following algorithm.

Algorithm 1 Generating Completely Dependent Random Variable Using $U(0, 1)$

```
Require: $X$ such that $X_i$ follows $U(0,1)$ for all $i \in 1, 2, ...n$
  for $0 \leq i \leq m^2 - 1$ do
    if $x \leq \frac{i+1}{m^2}$ then
      $c \leftarrow m(\frac{i}{m} - [\frac{i}{m}])$
      $r \leftarrow [\frac{i}{m}]$
      $i_2 \leftarrow mc + r$
      $k \leftarrow i_2 - i$
      return $x + \frac{k}{m^2}$
    end if
  end for
```

Chatterjee Correlation Coefficient

Let (X, Y) denote a pair of random variables where Y is not constant. Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be i.i.d pairs following same distribution as (X, Y), $n \geq 2$. We rearrange the data as

$$(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), \dots, (X_{(n)}, Y_{(n)})$$

by ordering the X_i's in lexicographic order, breaking any ties uniformly at random. Then, the cross correlation coefficient is given by

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

Here, r_i denotes the rank of each element Y_i and l_i denotes the number of j such that $Y_{(j)} \geq Y_{(i)}$. In case of no ties among Y_i's, $l_1, l_2, \dots, l_n$ is a permutation of $1, 2, \dots, n$, so the coefficient becomes

$$\xi(X, Y) = 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}$$

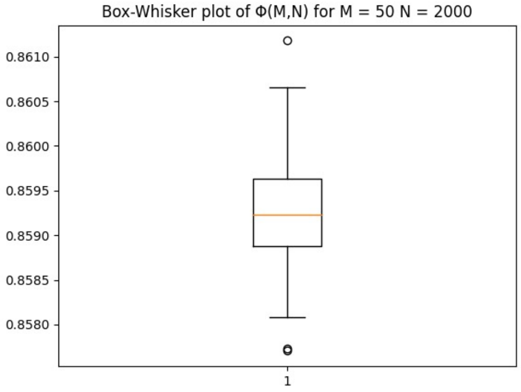
Algorithm 2 Chatterjee Correlation Coefficient

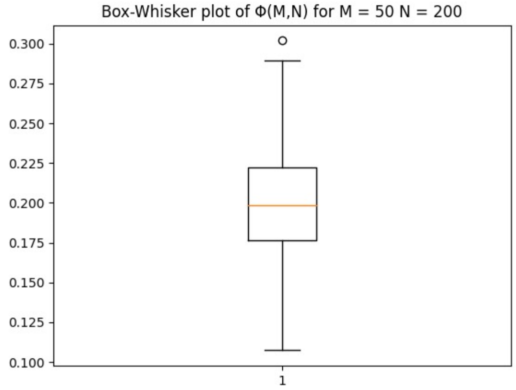
```
Require: (X,Y) $\in \mathbb{R}^n \times \mathbb{R}^n$
  Order $X_i$'s as $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)}$
  if $X_{(i)} < X_{(j)} \forall i < j$ and $Y_{(i)} \neq Y_{(j)} \forall i \neq j$ then
    Calculate $r_i = \{\#j : Y_{(j)} \leq Y_{(i)}\}$
    return
        $$\xi(X, Y) = 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}$$
  else if $X_{(j)} = X_{(j+1)} = \cdots = X_{(k)}$ for any $1 \leq j < k \leq n$ then
    Assign $m \in \{j, j+1, \dots, k\}$ to $X_{(i)}$, $j \leq i \leq k$, so that $P(X_{(i)} = m) = \frac{1}{k-j+1}$
    Calculate $r_i = \{\#j : Y_{(j)} \leq Y_{(i)}\}$
    Calculate $l_i = \{\#j : Y_{(j)} \geq Y_{(i)}\}$
    return
        $$\xi_n(X, Y) = 1 - \frac{n \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \sum_{i=1}^{n} l_i(n - l_i)}$$
  end if
```

Observations

We have carried out lot of simulations for different values of sample size(N), value of m, and here are some of the noteworthy observations.

1. Note that, for fixed m and n, the Chatterjee Correlation Coefficient between the random variables X and Y is a function of X, Y and hence is a real random variable whose range is a subset of [-1, 1]. Fixing m and n, for moderately large values of n, we can observe that the range of this statistic is quite small, so we can approximately consider the statistic as a degenerate random variable. The point at which it is degenerate however is dependent on both m and n, (as $X \sim U(0, 1)$ and Y is completely determined by m and X) hence we can treat this point as a function of m and n. We will denote this by $\xi(m, n)$, and the random variable by $\phi(m, n)$.





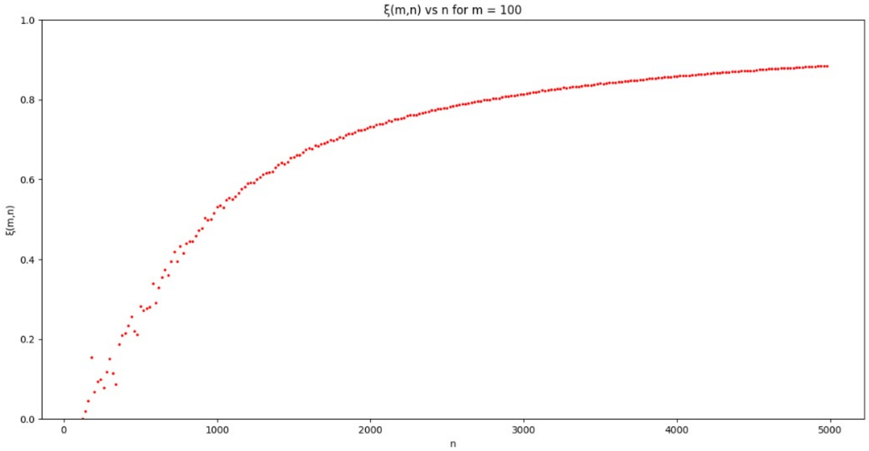


Figure 1: $\xi(m, n)$ vs. n for $m = 100$

2. For fixed M, as N $\rightarrow \infty$, $\xi(M, N)\rightarrow$ 1. In other words, since M determines the amount of complication in the bijection, we can say that if X and Y are related by any shuffle of M(might be a very complicated function), with a sufficiently large sample we can get cross-correlation arbitrarily close to 1

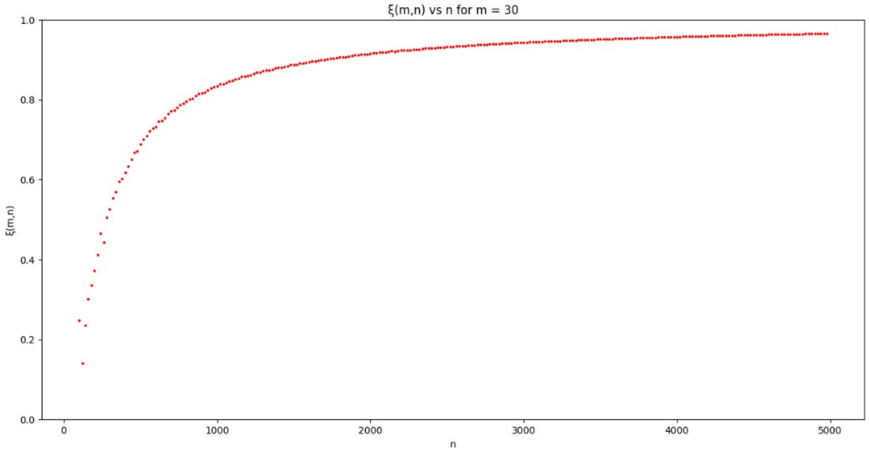


Figure 2: $\xi(m,n)$ vs. n for $m = 30$

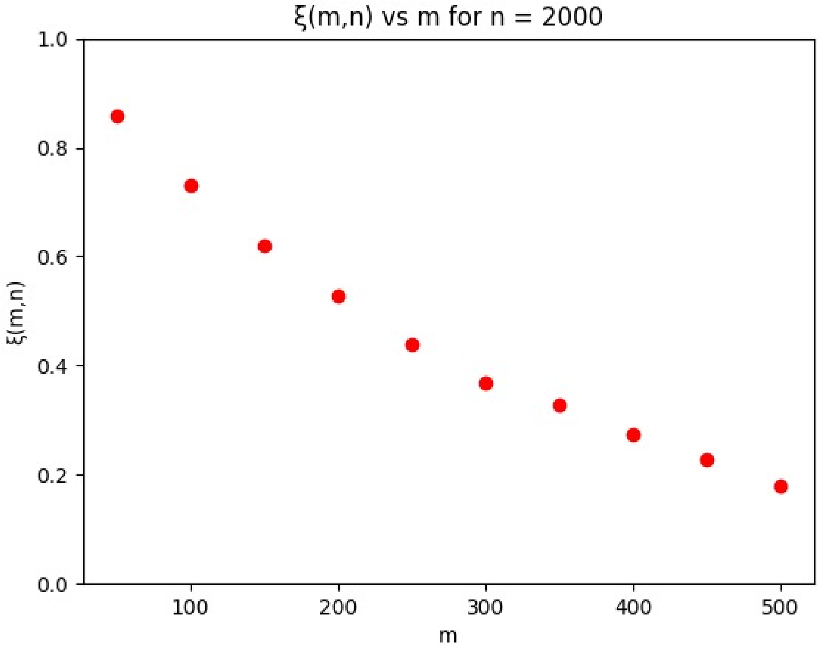


Figure 3: $\xi(m,n)$ vs m for $n = 2000$

3. For Fixed N, as M $\rightarrow \infty$, $\xi(M,N)\rightarrow 0$. In other words, if we keep the sample size fixed, then by increasing the shuffle of M, we can arbitrarily decrease the cross-correlation to 0.

4. This suggests for a fixed C in (0,1), there is a relation connecting all M and N such $\xi(M,N)$=C. This relation can be interpreted as how large a sample(N) should be taken such that if the value of M is less than or equal to some M(which varies), we have a cross-correlation greater than or equal to C. We have simulated for C = 0.9 and the values of M, N which produced this value roughly lie in a straight line.

All these observations are however from simulation, since the random variable $\phi(M,N)$ is a mathematically complicated quantity. Although most of the observations are intuitively correct, we try to give some mathematical justification of these observations in the next section.

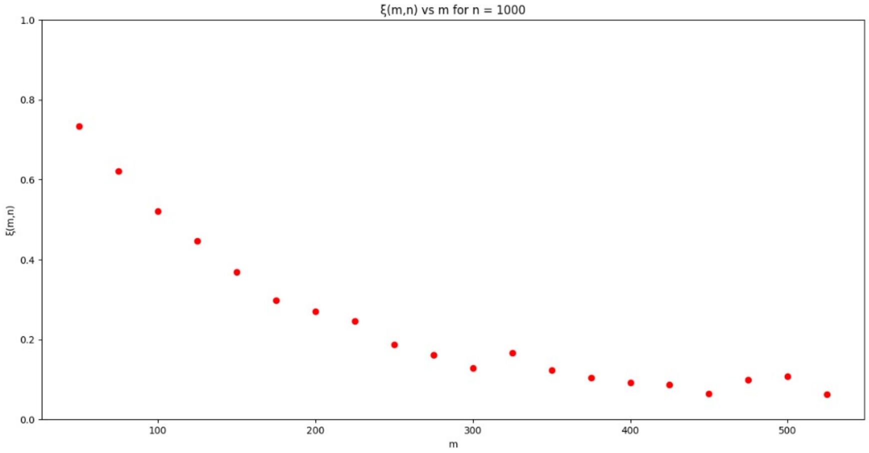


Figure 4: $\xi(m,n)$ vs m for $n = 1000$

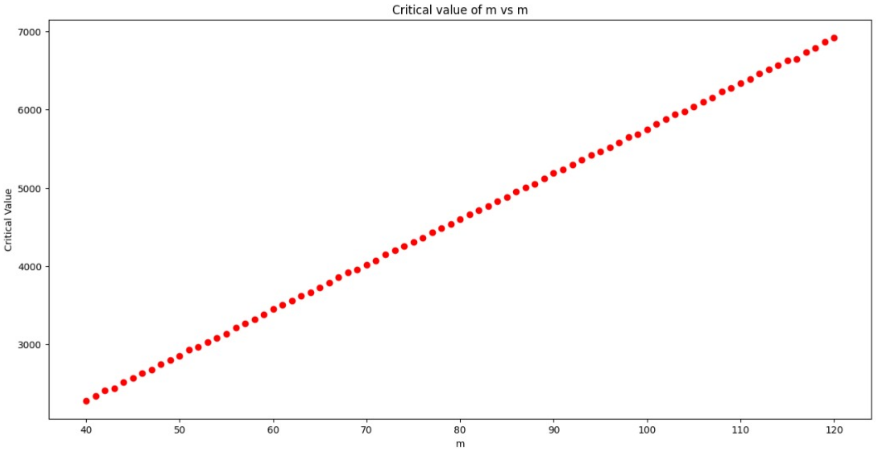


Figure 5: critical n vs m when $\xi(m,n) = 0.90$

Explanations

We observed that if we increased the shuffle of M, keeping sample size fixed the cross-correlation decreases to 0. On the other hand, keeping the shuffle of M fixed, if we increase the sample size, the cross-correlation increases to 1. Here we will try to give a possible explanation for this behaviour.

We have a result in probability that if $X_1, X_2, ...X_n$ are iid. Unif(0,1), then

$$E[X_{(i)}] = \frac{i}{n+1}$$

for all $i \in 1, 2, ...n$

Now we have simulated and observed for large N, that Substituting $X_i = \frac{i}{n+1}$ gives nearly the same value as $\psi(m,n)$. In other words,

$$E[CCC((X_1, X_2, ..., X_n), (f_M(X_1), f_M(X_2), ..., f_M(X_n)))]$$

is nearly equal to

$$CCC(E(X_{(1)}), E(X_{(1)}), ..., E(X_{(n)})), (f_M(E(X_{(1)})), f_M(E(X_{(2)})), ..., f_M(E(X_{(n)}))))$$

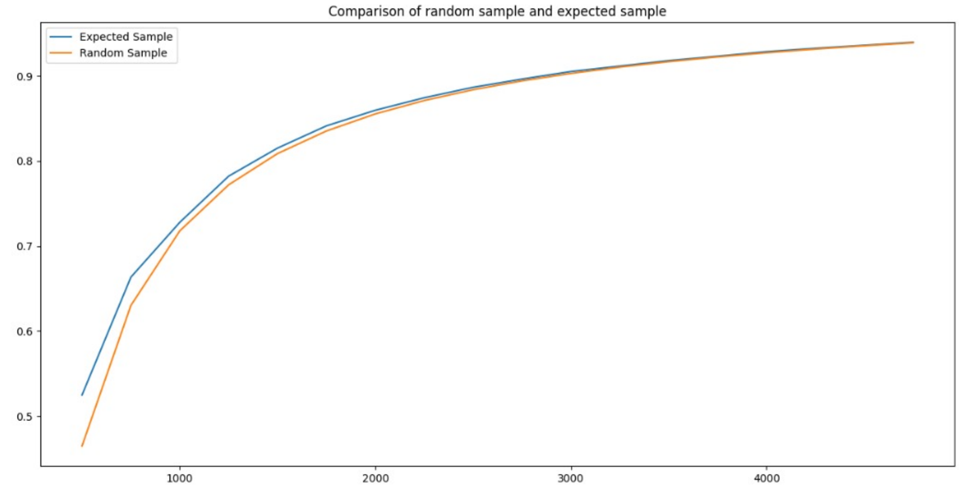


Figure 6: Comparison of random sample and expected sample

That is a fairly non-trivial fact itself, and we have no rigorous justification of that, but we have seen this holds by simulation. Assuming this, we can prove the result for large n.

Asymptotic behaviour with N

Note that, for a large sample size, $r_i =$ No. of y_j s.t $y_j \leq y_i \approx \pi(i)$. Now if $i = m(j_i - 1) + k_i$, then $\pi(i) = m(k_i - 1) + j_i$. Clearly, $r_{i+1} - r_i = m(k_{i+1} - k_i) + (j_{i+1} - j_i) = m - (m^2 - 1)(j_{i+1} - ji)$. Now note, $j_i = \lfloor (i-1)/m \rfloor + 1$. Hence we have $|r_{i+1} - r_i| = |(m^2-1)(\lfloor \frac{i}{m} \rfloor - \lfloor \frac{i-1}{m} \rfloor) - m|$ For a fixed m, we have $|r_{i+1} - r_i| = |(m^2-1)(\lfloor \frac{i}{m} \rfloor - \lfloor \frac{i-1}{m} \rfloor) - m| < (m^2-1)|(\lfloor \frac{i}{m} \rfloor - \lfloor \frac{i-1}{m} \rfloor)| + m < A(m)$, where A(m) is a constant depending only on m. [As, $b - a - 1 < \lfloor a \rfloor - \lfloor b \rfloor < b - a + 1$] Hence, $\sum_{i=1}^{n-1} |r_{i+1} - r_i| < n * A(m)$, which shows that as $n \to \infty$, $\frac{\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2-1} \to 0$ and hence, the correlation goes to 1.

Asymptotic behaviour with M

First, note that by increasing m, we can make $C_\epsilon(u, v)$ arbitrarily close to $\prod(u, v)$. Now suppose $(X, Y) \sim C_\epsilon(U, V)$ and $(X, Z) \sim \prod(U, V)$. Now, the joint distribution of (X, Y) is abitrarily close to the joint distribution of (X, Z). Now we claim, $|\text{Corr } (X, Y)| \approx |\text{Corr } (X, Z)|$ and as Corr $(X, Z) = 0$, we can say that, by increasing m, the cross-correlation decreases to 0.

Conclusion

We conclude this project by claiming that it has delved into the intricate relationship between Nelsen's copula theorem and cross-correlation analysis. Through our exploration, we have gained a deeper understanding of how a cross-correlation coefficient can detect a function (in our case, shuffles of M) with increase in sample size. Furthermore, our investigation into copula and shuffles of m has highlighted the fact that with a fixed sample size, nothing can be said about the independence of two random variables, only by knowing their joint distribution. In summary, this project explores cross-correlation for shuffles of M and tries to explain how Chatterjee Correlation Coefficient captures Nelsen's theorem.

Acknowledgements

We would like to express our sincere gratitude to all of the dedicated team members for their collaborative efforts and commitment in successfully completing this statistics project on **Exploring Cross Correlation Using Shuffles of M**. Each member's unique skills and contributions played a crucial role in the development and implementation of the algorithm. Additionally, our heartfelt thanks go to **Prof. Arnab Chakraborty** for his guidance and valuable insights throughout the project. This endeavor has been a rewarding learning experience, and we appreciate the support received from our peers , our parents and the academic community of ISI Kolkata.

References

1. S.Chatterjee,2021: *A new coefficient of correlation*, *Journal of Amer. Stat. Assoc.*

2. Roger B. Nelsen: *Introduction to Copulas*