Cross Correlation And Nelsen's Theorem

Arikith Roy Chowdhury

Under Supervision of: Prof. Arnab Chakraborty

April 2024

Introduction

Roger B Nelsen proved that given two independent random variables we can construct two random variables whose joint distributions can be made arbitrarily close to the joint distribution of two independent random variables with same individual probability distributions and yet one is a measurable function of the other. The aim of this project is to explore whether the well known Cross Correlation is able to detect the dependence of such pair of random variables which are "very close" to independent pairs. To construct such pairs we will be using a proof that our batch-mate Himadri Mandal came up with.

Theorem

Let X,Y be two independent continuous random variables with distributions F_X, F_Y. Given $\epsilon > 0$, $\exists$ random variables U,V with distributions F_U, F_V respectively such that
(i)$F_X = F_U,\ F_Y = F_V$
(ii)$sup\ |F_{X,Y}(a,b) - F_{U,V}(a,b)| < \epsilon, (a,b) \in \mathbb{R}^2$
(iii) U is a measurable function of V

The detailed proof of the theorem is attached with this report. For simulations we will be working with continuous random variables X, Y with range [0,1]. To attain this we instead work with $\frac{c}{1+e^{X_0}}$ where X_0 comes from some standard distribution and c is a constant chosen suitably to make the range [0,1]. Our random variable generating function is

$$\boxed{g_n(y) = F_X^{-1}\left(F_Y(y) + \left((i-j)\cdot\left(\frac{n-1}{n^2}\right)\right)\right)}$$

Here is i, j is as defined in the proof. $g_n(Y), Y$ serves as U, V respectively satisfying the above mentioned properties. .

Simulation Results

n= proximity parameter, N= Sample Size, $\xi_{N,n}$= Cross Correlation

1. $X = \frac{1}{1+e^{X_0}}, Y = \frac{1}{1+e^{Y_0}};\ X_0 \sim N(0,1),\ Y_0 \sim N(0,1),\ n = 500$

First we verify that our function indeed satisfies the desired properties, i.e property (i) and (ii) of the theorem. Figure 1 shows the plots of the empirical CDFs of X and U for N=5000 and n=500

We can clearly see that their empirical CDFs are very close and almost indistinguishable and as $N \to \infty$ the ecdfs converge to F_X by Glivenko-Cantelli theorem.

Figure 2 shows the plots of empirical joint CDFs $F_{X,Y}$ and $F_{U,V}$ in red and green respectively when n=500, N=1500. The 3D plot is available here. Here also their proximity is as it was required.

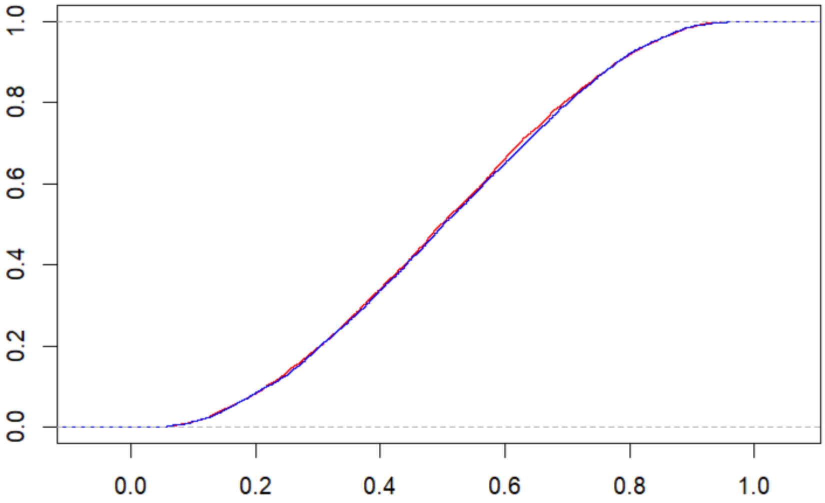


Figure 1: Empirical CDFs of X, U when N=5000

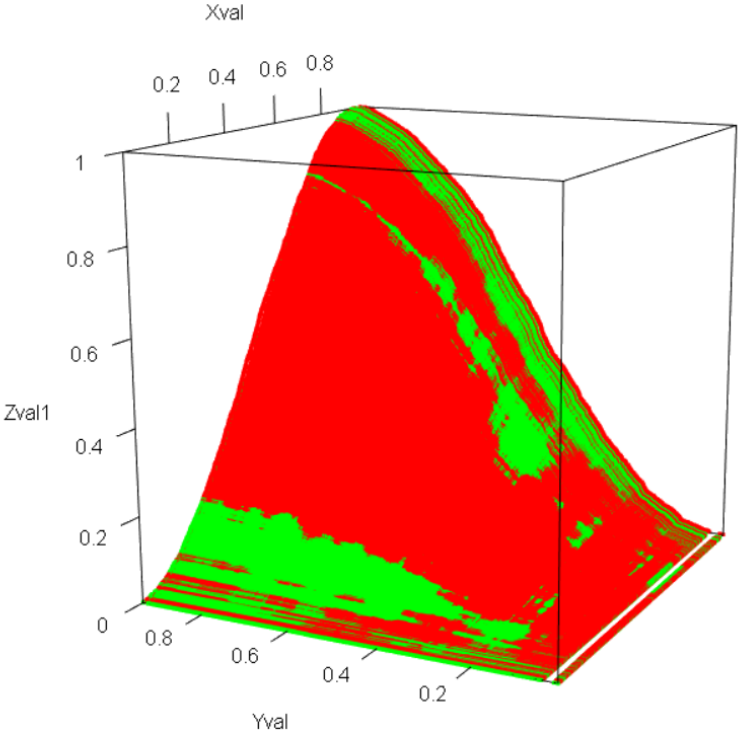


Figure 2: Empirical Joint CDFs

We now proceed to compute the cross correlation values of (X, Y) and (U, V) for different values of N with n=500. Some values of ξ_N are given in the table below.

Sample Size(N)	$\xi_N(X, Y)$	$\xi_N(U, V)$
100	0.0219	-0.0447
200	-0.0501	0.0561
500	0.0376	0.0412
1000	0.0080	0.0446
1500	0.0200	0.0446
2000	-0.0149	0.2043
4000	0.0041	0.4416
7000	-0.0053	0.6356
10000	-0.0100	0.7301
20000	-0.0009	0.8575
50000	-0.0025	0.9412
100000	0.0021	0.9703

Table 1: Computed ξ_N Values

Observing the ξ_N values it turns out that for sample size up to 2000-3000 cross correlation coefficient is relatively small, But as the sample size increases, ξ_N values also increases and reaches 0.97 for $N = 100000$. Figure 3, 4 shows ξ_N vs N plots for both pairs.

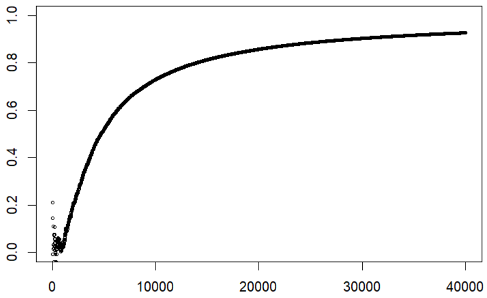


Figure 3: $\xi_N(U, V)$ vs N

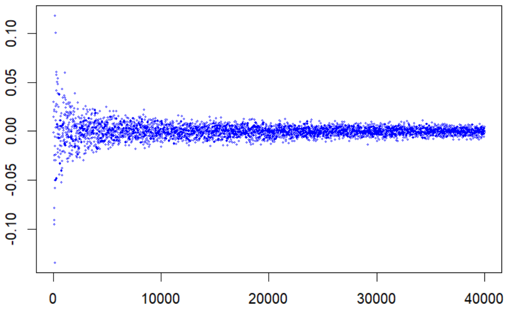
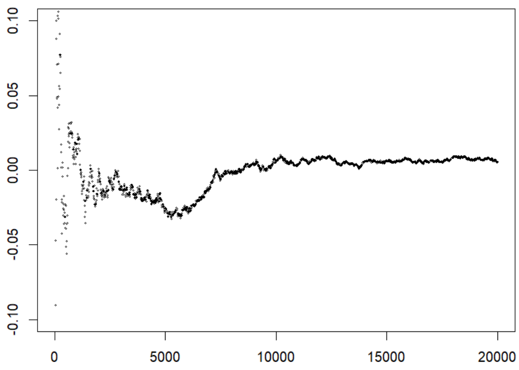


Figure 4: $\xi_N(X, Y)$ vs N

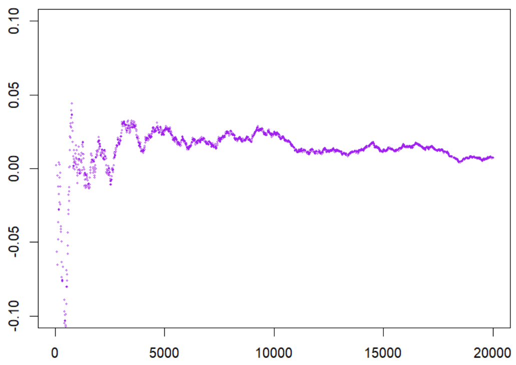
- Clearly as $N \to \infty$ $\xi_N \to 1$. Hence cross correlation is indeed able to detect deterministic dependence asymptotically. Another interesting thing is that in figure 3 the plot appears to be a nice mathematical curve for all but some smaller values of N, where it oscillates quite a bit.

- For (X, Y), ξ_N values are always very small and as N increases they come closer to 0 proclaiming their independence.

At this point one may want to see how the classical coefficients of correlation(Spearman's ρ, Kendall's τ and Pearson's r) behave in this given scenario. Well unfortunately all of them fail terribly to detect the deterministic dependence of U and V. The plots are shown below (n is fixed to be 500).



$r(U, V)$ vs N

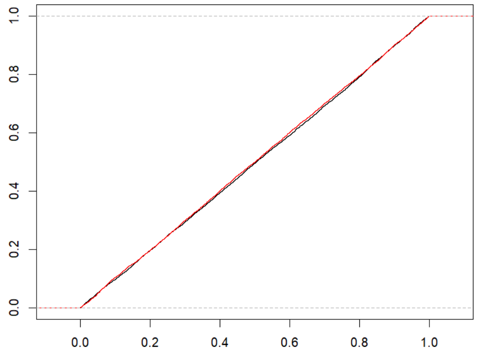


$\tau(U, V)$ vs N

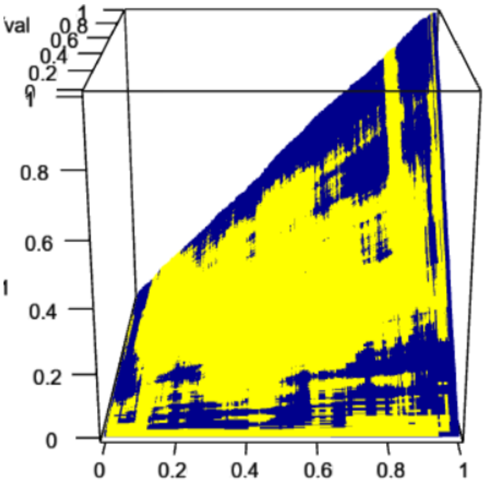
Note: In the original published paper of cross correlation it was proved that $\lim_{n\to\infty} \xi_n(X, Y)$ is 0 if and only if X and Y are independent, and it is 1 if and only if there is a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$ almost surely. Since our g_n is a measurable function, for any choice of random variables X,Y(satisfying the required conditions) the value of $\xi_N(U, V)$ goes to 1 as sample size $N \to \infty$.

2. $X \sim Unif(0,1),\ Y \sim Unif(0,1)$

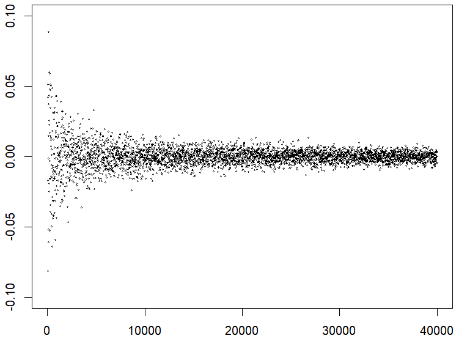
In this case there is no need to do any transformation as the random variables are already absolutely continuous in the range [0,1]. Simulation results are similar and plots are shown below.



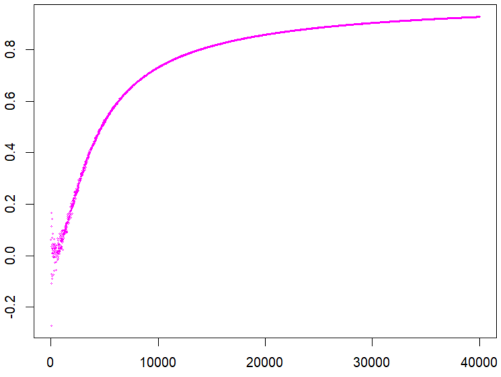
Ecdfs of X,Y for N=10000



E.Joint CDFs, N=1000



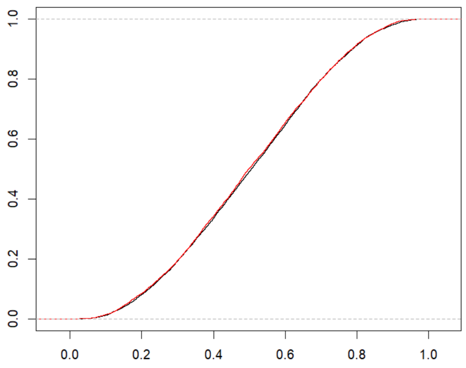
$\xi_N(X,Y)$ vs N



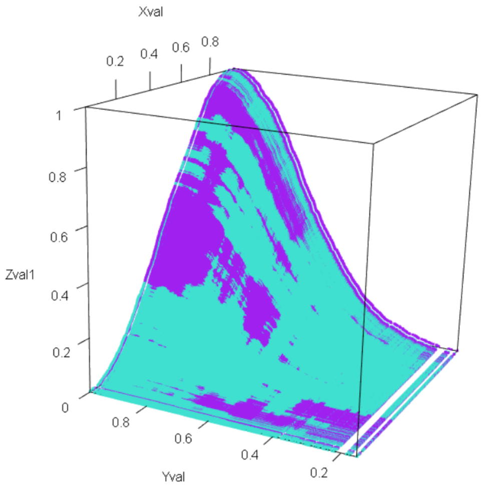
$\xi_N(U,V)$ vs N

3. $X = \frac{1}{1+e^{X_0}}, Y = \frac{2}{1+e^{Y_0}};\ X \sim N(0,1),\ Y \sim \Gamma(3,4),\ n = 500$

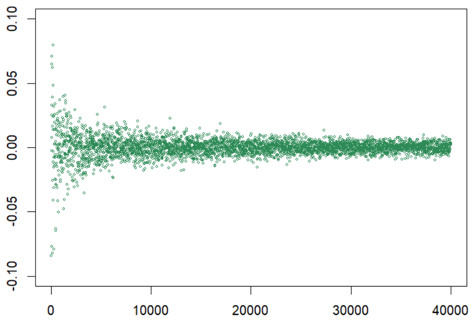
Firstly it is important to note that since Γ distribution has support $[0,\infty)$, we have taken Y to be $\frac{2}{1+e^{Y_0}}$ so that F_Y has support [0,1]. Recall that our function g_n is valid only for distributions with support [0,1].



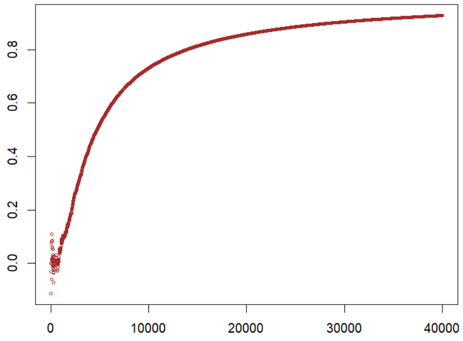
Ecdfs of U,X for N=8000



E.Joint CDFs, N=1000



$\xi_N(X,Y)$ vs N



$\xi_N(U,V)$ vs N

Analyzing $\xi_{N,n}(U,V)$

In the following 2 sections we will try to see how $\xi_{N,n}$ behaves when one of the parameters N or n is fixed and other is varied. We will also try to fit functions on the plots and try to find the parameters involved.

1. Varying n with N fixed

In all the previous cases the value of the proximity parameter n was fixed. We will now see the behaviour of $\xi_n(U,V)$ with varying n for different fixed values of N. The following figure shows the scatter plot for 3 different values of N (mentioned in label). Setup is same as in case 3.

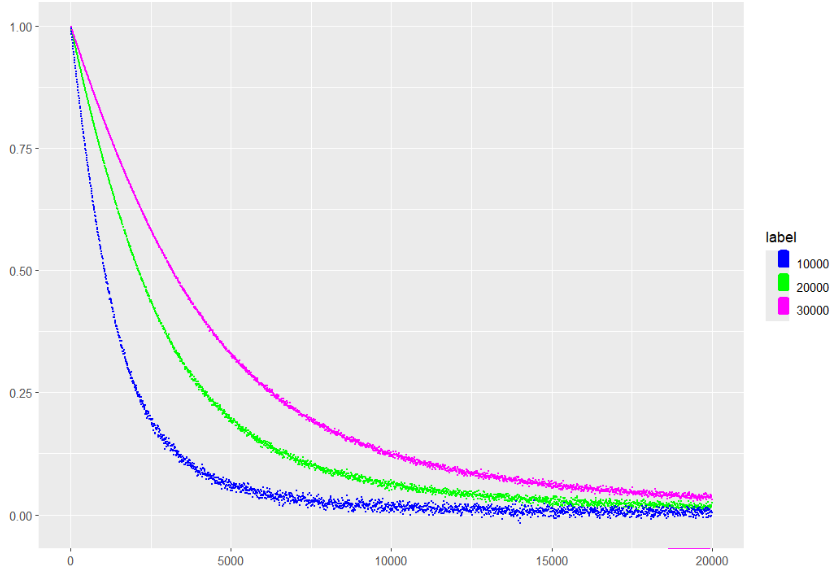


Figure 5: $\xi_n(U,V)$ *vs* n plots

- For a fixed sample size, as the value of n increases, random variables U, V become more and more 'closer' to X, Y making dependence harder to detect and naturally that results is decrease of ξ_n values. As seen clearly in Figure 5, ξ_n values decreases to 0 as $n \uparrow$, causing cross correlation to fail eventually.
- On the other hand, on the increase of sample size, ξ_n values increase which is evident from the preceding plot.

But what function does the ξ_n follow here? And what are the parameters? To answer these we now make a combined $\xi_n(U,V)\ vs\ n$ plots for all the 3 cases(Figure 6), fixing N=10000. And to our surprise these plots superimpose each other perfectly which shows that the certain function being followed is **independent of the choice of the random variables**!!! We will now try to trace out this function that ξ_n follows as n varies for a fixed N. At a first glance the curves appear to follow an exponential decay. Let us try to fit a function of the form e^{-cx}.

$$\xi_n \sim e^{-cn} \implies -log\xi_n \sim cn$$

Figure 7 shows $-log\xi_n$ vs n plot for case 1. For very large values of n (w.r.t N) the -log ξ_n vs n plot becomes very much scattered(since in that case the behaviour of the correlation coefficient is very similar to $\xi_n(X,Y)$). Therefore we restrict ourselves to smaller values of n for performing the regression. Clearly the relationship is almost linear. We fit a straight line and obtain the slope c=0.0006198419. To verify the model we plot e^{-cn} and as seen in figure 8 the fit is **excellent**! So we hypothesize $\boxed{\xi_n(U,V) = e^{-cn}}$ for a fixed N.(c depends on N only). [1]

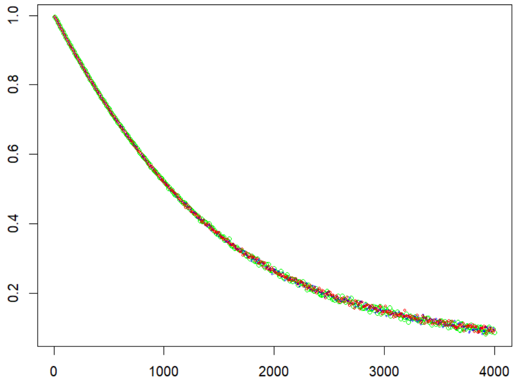


Figure 6: Combined $\xi_n(U,V)\ vs\ n$ plot

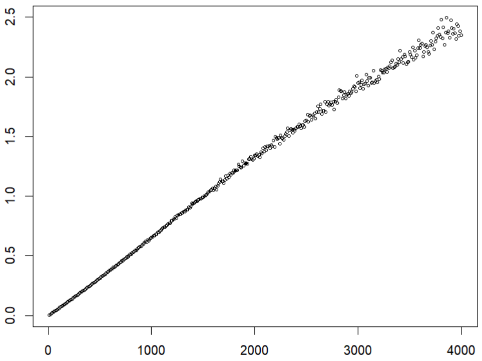


Figure 7: $-log\xi_n$ vs n

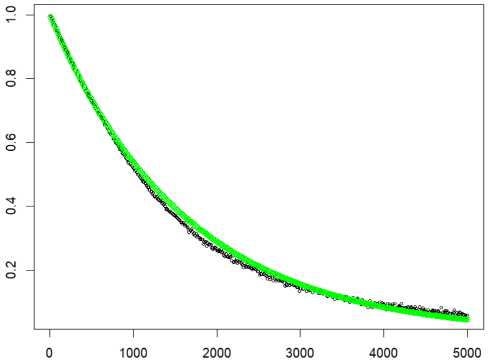
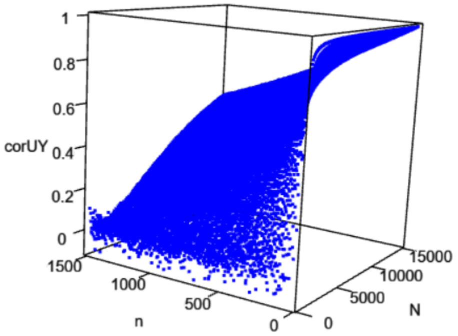


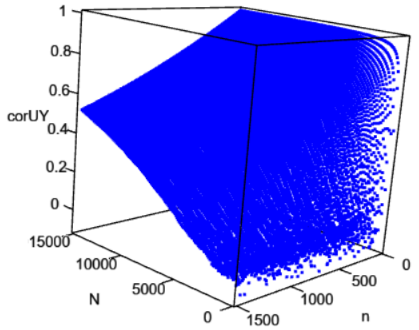
Figure 8: Fitted function

[1] the superimposition is because ξ is invariant under any monotone transform

For a clear visualization of how $\xi_{N,n}(U, V)$, n and N vary as the others change we make a 3D plot (for the preceding case). The 3D plot can be seen here. 2D views are given below.



View 1



View 2

2. Varying N with n fixed

Looking back to the $\xi_N(U, V)$ vs N plot, it appears that they follow a certain function that possibly does not depend on the choice of the random variables. So we make the combined plot of all the 3 cases and it turns out that our guess is indeed **TRUE!**

Figure 9 shows the combined $\xi_N(U, V)$ plot of all the 3 previous cases with n=500 fixed. As can be seen, they are superimposing each other perfectly! Hence we make the hypothesis that the function it follows is **independent of the choice of the random variables**. It depends only on N. That's truly **amazing!** In the following section we trace out the function.

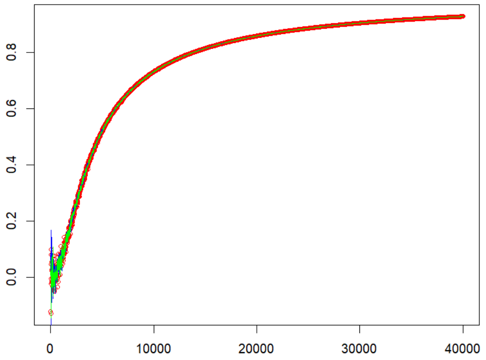


Figure 9: Combined $\xi_N(U, V)$ vs N

Looking at this plot, the very first function we can think of is something of the form $\frac{1}{1+c/x}$, with some unknown coefficient c. As x goes to 0 the function goes to 1 and it is 0 at 0. Another such function can be $\frac{1}{1+ce^{-x}}$. There can also be some linear function of x in places of x in the preceding functions. Let us assume that the function is of the form $\frac{1}{1+f(x)^{-1}}$. We will first try to fit a straight line to $f(x)$. We have

$$\xi_N(U,V) \sim (1+f(N)^{-1})^{-1} \implies f(N) \sim \frac{1}{\xi_N(U,V)^{-1}-1}$$

We plot $(\xi_N(U,V)^{-1}-1)^{-1}$ vs N. And here we go!!! The plot shows almost perfect linear relationship (Figure 10). This is very fascinating.

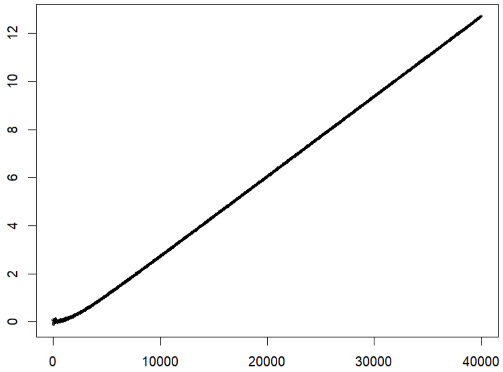


Figure 10: $(\xi_N(U,V)^{-1}-1)^{-1} vs N$

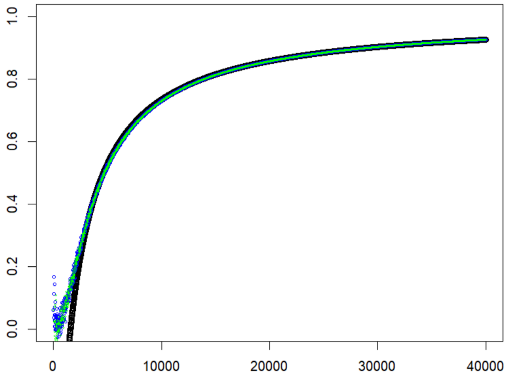


Figure 11: Fitting the function

We fit a straight line on it and the slope(m) and intercept(c) turns out to be 0.000328 and -0.5206 respectively. Hence we arrive at the following equation.

$$\boxed{\xi_N(U,V) = (1+(mN+c)^{-1})^{-1}}$$

. Surely the values of m and c depend on n(as different values of n render different plots). To verify this equation we plot the function to the combined plot.

As seen in figure 11, the function we obtained renders a perfect fit. This proves its validity. Now the following question arise.

How exactly does the values of m and c depend on n? Can we find m,c as functions of n? [2]

[2] this function also possibly depends on the construction of g_n

Convergence Result

The original paper ”A NEW COEFFICIENT OF CORRELATION”, does not provide any asymptotic theory for dependent random variables, however the following theorem is proved there.

Theorem: Suppose that X and Y are independent and Y is continuous. Then $\sqrt{N}\xi_N(X,Y) \to$ N(0, 2/5) in distribution as $N \to \infty$.

Now for a fixed value of N, as $n \to \infty$ the joint distribution of (U,V) converges to the joint distribution of (X,Y), so for very large n we may expect similar behaviour of $\sqrt{N}\xi_N(U,V)$ as stated in the preceding theorem.

In order to simulate this we take a very large value of n=10^6 and sample size N=1000. And surprisingly the following plots shows a excellent fit with $N(0,2/5)$. Histograms of 5000 an 10000 simulations are shown below. [3] [4]

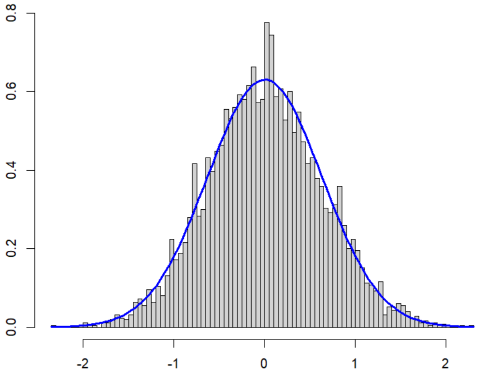


Figure 12: Histogram of 5000 simulations of $\sqrt{N}\xi_{N,n}(U,V)$ when N=1000, n=10^6

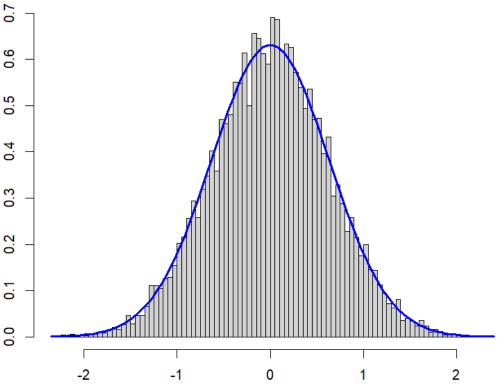


Figure 13: Histogram of 10000 simulations of $\sqrt{N}\xi_{N,n}(U,V)$ when N=1000, n=10^6

[3] the plots are done for case 1

[4] the blue curve is the PDF of N(0,2/5)

The simulation results prompts us to find the answers to the following questions.

Points To Ponder

For a fixed sample size N, we can choose sufficiently large n such that the behaviour of $\xi_{N,n}(U,V)$ is similar to that of $\xi_N(X,Y)$ and the distribution of $\sqrt{N}\xi_{N,n}(U,V) \approx N(0,2/5)$. However if we now increase N keeping n fixed the approximation will fail after a certain value. Then can we find the conditions on N and n under which we can very well use this approximation ?

However based on the experiments and simulations done before, what we CAN say at this point is the following.

Given any finite data of 2 random variables we can never guarantee their independence . (Neither cross correlation nor any of the standard correlation is able to do this and to this date we don't have any such coefficient of correlation which can do that.)

Summary

1. We used a function g_n to generate pairs of random variables ($g_n(Y)$,Y) from independent pairs (X, Y) which serve as our desired random variables of Nelsen's theorem and carry out simulations.

2. Asymptotically the value of cross correlation goes to 1 iff one random variable is a measurable function of the other and goes to 0 iff they are independent.

3. The way we constructed our random variable pairs, satisfy the conditions of Nelsen's theorem, however since $U = g_n(V)$ and g_n being measurable, cross correlation is indeed able to detect the dependence asymptotically regardless of how close their joint CDFs maybe to the independent X,Y.

4. For a given sample size N we can choose sufficiently large n so that the cross correlation is not able to the dependence(i.e the correlation values are very small). And in that case the distribution of $\sqrt{N}\xi_{N,n}$ is very well approximated by N(0,2/5).

5. Given any finite data of 2 random variables we can never guarantee their independence.

6. For a fixed N we get $\xi_n(U, V) \approx e^{-cn}$, where c depends only on N and is independent of the specific choices of the random variables. This does not hold from a certain large value of n onwards when the correlation values is very close to 0.

7. For a fixed n we get $\xi_N(U, V) \approx \frac{1}{1+(mN+c)^{-1}}$ where m and c depends only on n and is independent of the specific choice of the random variables. This does not hold for some initial values of N when the fluctuations are very frequent.

8. The function g_n being very simple and easy to compute and can be used as an efficient tool to determine how good some specific measure of independence is.