Applications of Chatterjee's Correlation in MCMC

Vivek Kumar Singh

BS/MTH/190988

Supervisor: Dr. Dootika Vats

March 4, 2022

Contents

Chapter 1

About the new coefficient of correlation paper

1.1 Problems with correlation coefficients

1. The three most popular classical measures of statistical association are Pearson's correlation coefficient, Spearman's ρ, and Kendall's τ.

2. They are very good at detecting linear or monotone relations between variables, and have well developed statistical theories. But they fail to detect non-monotone associations (e.g. periodic relations).

3. Many coefficients have been proposed in the past to address these issues. But most of them have these two problems.

   (a) One would like the coefficient to be close to its maximum value iff one variable is close to being a noiseless function of the other. For example in Pearson coefficient, it is 1 iff one variable is close to being a linear function of the other. They are 1 when it one variable is a noiseless function of the other but the converse doesn't hold.

   (b) They do not have simple asymptotic theories under the hypothesis of independence that facilitate the quick computation of p-values for testing independence. We have to rely on expensive permutation tests or bootstrap to test for independence.

4. The Chatterjee correlation coefficient (not the official name) presented in the next section addresses the above two issues.

1.2 Chatterjee Correlation Coefficient

This coefficient is (a) as simple as the classical ones, (b) is a consistent estimator of some measure of dependence which is 0 iff the variables are independent, and 1 iff one is a measurable function of the other, and (c) has a simple asymptotic theory under the hypothesis of independence, like the classical coefficients.

Let (X, Y) be a pair of random variables, where Y is not a constant (for our purposes, both X and Y are continuous). Let $\{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. pairs following the same distribution as (X, Y).

1. The case when $X_i's$ and $Y_i's$ have no ties. Rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} < \cdots < X_{(n)}$. Let r_i be the rank of $Y_{(i)}$, i.e. the number of j such that $Y_{(j)} \leq Y_{(i)}$. Then the correlation coefficient ξ_n is defined to be
$$\xi_n(X, Y) := 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}$$
.

2. In the case of ties. If there are ties in $X_i's$, choose an increasing arrangement as follows and break ties uniformly at random. Let r_i defined as above, and define l_i to be the number of j such that $Y_{(j)} \geq Y_{(i)}$. Define
$$\xi_n(X, Y) := 1 - \frac{n\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2\sum_{i=1}^{n-1} l_i(n - l_i)}$$
. When there are no ties among the $Y_i's$, $l_1, \ldots, l_n$ is just a permutation of $1, \ldots, n$ and the denominator is just $n(n^2 - 1)/3$, which reduces to the definition in the no ties case.

Theorem 1.1. If Y is not almost surely a constant, then as $n \to \infty$, $\xi_n(X, Y)$ converges almost surely to the deterministic limit

$$\xi(X, Y) := \frac{\int Var(\mathbb{E}(1_{\{Y \geq t\}}|X))d\mu(t)}{\int Var(1_{\{Y \geq t\}})d\mu(t)}$$

where μ is the pdf of Y. This limit belongs to the interval $[0, 1]$. It is 0 iff X and Y are independent, and it is 1 iff there is a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$ almost surely.

1.2.1 Questions relevant to our purpose and some TO-DOs

1. Develop an intuitive understanding of Theorem 1.1 [Read the proof given in Section 8 and [10] of the main paper].

2. How good is the estimator $\xi_n(X, Y)$ if our data is not i.i.d.? [Simulate some basic examples. Do for normal distribution (both for i.i.d. and MCMC data)]

3. If $X_1, \ldots, X_n$ is a MC, is $\xi(X_k, X_{k+j})$ independent of k as seen in the Pearson correlation, i.e. same for all k?

4. If $X_1, \ldots, X_n$ is a MC, $\xi(X_k, X_{k+j}) \to 0$ as $j \to \infty$? It seems true intuitively as as $\xi(X, Y) = 0$ iff X, Y are independent, and in this case, as X_k, X_{k+j} are independent as $j \to \infty$ [Prove that the independence happens in the limiting case and done!]

5. For the above MC, is $\xi(X_k, X_{k+j})$ similar to $Corr(X_k, X_{k+j})$ for similar j? [Simulate some basic examples, AR(1), AR(2)]

1.2.2 Some remarks till now

1. ξ_n is not symmetric in X, Y. This is intentional and useful as we might want to study if Y is a measurable function of X, or X is a measurable function of Y.
   To get a symmetric coefficient, it suffices to consider $max(\xi_n(X, Y), \xi_n(Y, X))$. This estimator by Theorem 1.1 will converge to the maximum of respective limits. This will be equal to 0 iff X, Y are independent, and 1 iff X is a measurable function of Y or Y is a measurable function of X.

2. $\xi(X, Y) \in [0, 1]$
   Proof: We know that for any random variables X and Y,

$$Var(X) = Var(\mathbb{E}(X|Y)) + \mathbb{E}(Var(X|Y))$$

$$\implies Var(1_{\{Y\geq t\}}) \geq Var(\mathbb{E}(1_{\{Y\geq t\}}|X))$$

Hence the integral will always be in [0, 1].

3. If X and Y are independent, then $\mathbb{E}(1_{\{Y\geq t\}})$ is a constant, and so $\xi(X,Y) = 0$
   If X is a measurable function of Y, then $\mathbb{E}(1_{\{Y\geq t\}}|X) = 1_{\{Y\geq t\}}$, and so $\xi(X,Y) = 1$.

4. If there are no ties among $Y_i's$, then the maximum possible value of $\xi_n(X,Y)$ is $\frac{n-2}{n+1}$, attained when $X_i = Y_i$ for all i. This is very small for small n. Say for $n = 20$, maximum value of $\xi_n(X,Y)$ is 0.86.

5. On the other hand, the smallest value of $\xi_n(X,Y)$ is $-1/2 + O(1/n)$, attained when top $n/2$ values of Y_i are placed alternately with the bottom $n/2$ values. This is strange as it suggests that the value can go way below 0, but theorem 1.1 states that in the limiting case, it is non-negative. The only interpretation of a negative value is that the data used for estimation is correlated (Our MCMC case).

Chapter 2

Some new results related to the Chatterjee correlation and Markov chains

Theorem 2.1. Let $X_1, X_2, \dots$ be a stationary, time homogeneous Markov chain with distribution π. We claim that

1. $Cov(X_k, X_{k+t})$ is independent of k.

   Proof. We know that

   $$Cov(X_k, X_{k+t}) = \mathbb{E}[X_k X_{k+t}] - \mathbb{E}[X_k]\mathbb{E}[X_{k+t}]$$

   Also, as X_n is a stationary markov chain, $\mathbb{E}[X_n] = \mu$, where μ is the mean of the distribution π. And,

   $$\mathbb{E}[X_k X_{k+t}] = \int \int xy f_{(X_k, X_{k+t})}(x, y) dx dy$$

   where $f_{(X_k, X_{k+t})}$ is the joint density of X_k and X_{k+t}. Now as the markov chain is stationary, this density is dependent only on t, i.e.

   $$f_{(X_k, X_{k+t})} = f_{(X_1, X_{1+t})}$$

   As both the terms of $Cov(X_k, X_{k+t})$ are independent of k, $Cov(X_k, X_{k+t})$ is independent of k.

   □

2. $\xi(X_k, X_{k+t})$ is independent of k.

Proof.

$$\xi_{(X_k,X_{k+t})} = \frac{\int Var(\mathbb{E}[1_{\{X_{k+u}\geq t\}}|X_k = x])d\pi(t)}{\int Var(1_{\{X_{k+u}\geq t\}})d\pi(t)} \tag{2.1}$$

We'll prove that both the numerator and the denominator are independent of k.
We can write

$$\begin{aligned} \mathbb{E}[1_{\{X_{k+u}\geq t\}}|X_k = x] &= Pr(X_{k+u} \geq t|X_k = x) &(2.2)\\ &= Pr(X_{k+u} \in [t,\infty)|X_k = x) &(2.3)\end{aligned}$$

and by time-homogeneity of our Markov chain

$$Pr(X_{k+u} \in [t,\infty)|X_k = x) = \int_t^\infty P^u(x, dy) \tag{2.4}$$

which is independent of k. And hence,

$$\int Var\left(\int_t^\infty P^u(x, dy)\right) d\pi(u) \tag{2.5}$$

is also independent of k.

Now for the denominator, we know by stationarity of our Markov chain that $X_n \sim \pi$, so for any function f, $f(X_n) \sim \pi'$, and therefore the variance will be same for all n.
Let $f_t : \mathbb{R} \to \mathbb{R}$ such that $f_t(X) = 1_{\{X\geq t\}}$.
We can write the denominator as

$$\int Var(f_t(X_{k+u}))d\mu(t)$$

where,

$$Var(f_t(X_{k+u})) = Var(f_t(X_1))$$

Therefore,

$$\int Var(f_t(X_{k+u}))d\mu(t)$$

is independent of both k and u.

As both the numerator and denominator are independent of k, we can conclude that $\xi(X_k, X_{k+u})$ is independent of k. □

3. $\xi(X_k, X_{k+t}) = \xi(X_{k+t}, X_k)$ for time reversal Markov chains.

   Proof. From (2), we know that the denominator of $\xi(X_k, X_{k+t})$ is independent of both k and t. So we only have to prove that the numerator is symmetric.
   Now for a time reversal Markov chain, we know that
   (Not sure if this is correct, need to prove/confirm)

   $$Pr(X_{n+k} \in A | X_n = x) = Pr(X_n \in A | X_{n+k} = x)$$

   So, by (2.4) we have,

   $$\begin{aligned} Pr(X_k \in [t, \infty) | X_{k+u} = x) &= Pr(X_{k+u} \in [t, \infty) | X_k = x) \\ &= \int_t^\infty P^u(x, dy) \end{aligned}$$

   Hence the numerator in both the cases will be equal to (2.5).
   As both the numerator and denominator are symmetric, we are done. $\square$

4. $\lim_{t \to \infty} \xi(X_1, X_{t+1}) = 0$ for an Ergodic Markov chain

   Proof. We have

   $$\xi(X_1, X_{t+1}) = \frac{\int Var(\mathbb{E}[1_{\{X_{t+1} \geq u\}} | X_1 = x]) d\pi(u)}{\int Var(1_{\{X_{t+1} \geq u\}}) d\pi(u)}$$

   The denominator is independent of t as proven in (2), so we only need to show that the numerator goes to 0 as $t \to \infty$.

   Lemma 2.2.

   Now, by (2.4), we know that

   $$\mathbb{E}[1_{\{X_{t+1} \geq u\}} | X_1 = x] = \int_u^\infty P^t(x, dy)$$

   For an Ergodic Markov chain, under the Total Variation Norm, we know that

   $$||P^t(x, \cdot) - F(\cdot)|| \to 0 \text{ as } t \to \infty$$

   $\square$

Chapter 3

Proof of symmetricity in a special case

Theorem 3.1. Let $X_1, \ldots, X_n$ be a time reversible Markov chain, then

$$\xi(X_n, X_{n+1}) = \xi(X_{n+1}, X_n)$$

Proof. As the Markov chain is time reversible, by Chapter 1, Section 5 of the Handbook of MCMC, the distribution of $(X_i, X_{i+1}, \ldots X_j)$ is same as $(X_j, X_{j-1}, \ldots X_i)$. In particular, distribution of (X_n, X_{n+1}) is same as (X_{n+1}, X_n). This means that for any $(x, y) \in \mathbb{R}^2$, we have

$$f_{(X_n, X_{n+1})}(x, y) = f_{(X_{n+1}, X_n)}(x, y)$$

We are only interested in the numerator of ξ as the denominator only depends on π as shown in Theorem 2.1.2.
We want to show that,

$$\int (Var(Pr(X_{n+1} \geq t|X_n)) - Var(Pr(X_n \geq t|X_{n+1})))d\pi(t) = 0$$

We claim that

$$Var(Pr(X_{n+1} \geq t|X_n)) - Var(Pr(X_n \geq t|X_{n+1}) = 0$$

Proof.

$$Var(Pr(X_{n+1} \geq t|X_n)) - Var(Pr(X_n \geq t|X_{n+1})$$
$$= Var(\int_t^\infty f_{(X_{n+1}|X_n=x)}(y)dy) - Var(\int_t^\infty f_{(X_n|X_{n+1}=x)}(y)dy)$$

$$= Var(\frac{1}{f_{X_n}(x)}\int_t^\infty f_{(X_n,X_{n+1})}(x,y)dy) - Var(\frac{1}{f_{X_{n+1}}(x)}\int_t^\infty f_{(X_{n+1},X_n)}(x,y)dy)$$

$$= Var(\frac{1}{f_{X_n}(x)}\int_t^\infty f_{(X_n,X_{n+1})}(x,y)dy) - Var(\frac{1}{f_{X_n}(x)}\int_t^\infty f_{(X_n,X_{n+1})}(x,y)dy)$$

$$= 0$$

$\square$

Using this claim, we get the desired result. $\square$

Chapter 4

TODOs and Questions to ask in the meeting

TODOs

1. For intuition of Theorem 1.1, read 1910.12327 and [10].

2. By Theorem 2.2, now we can think of estimating ξ using a single Markov chain. Need to see by simulation how the estimates are if our data is not iid.

Questions

1. How does reversibility imply that the distribution of $(X_i, \ldots, X_j)$ is same as $(X_j, \ldots, X_i)$?

2. If we leave out some of the indices while considering the joint distribution, will they still remain the same?