Applications of Chatterjee's Correlation in MCMC

Vivek Kumar Singh

BS/MTH/190988

Supervisor: Dr. Dootika Vats

April 14, 2022

Contents

1. Introduction

Markov chain Monte Carlo (MCMC) methods are a class of algorithms used for sampling from complicated probability distributions. They are often required for parameter estimation in the statistical models encountered in real-world applications.

If $X_1, X_2, \ldots$ is the Markov chain, then the lag-k autocorrelation is defined as

$$\gamma_k = \rho(X_1, X_{1+k}),$$

where ρ is the Pearson's correlation coefficent (See Section []). The autocorrelation function is used for assessing the quality of the Markov chain produced.

The Pearson's correlation is great at detecting monotone relations in data, but fails when applied to non-linear situations (say $Y = \sin(X)$)

Chatterjee proposed a new measure of dependence in [site paper] which is (a) as simple as the Pearson correlation, (b) is a consistent estimator of some measure of dependence which is 0 if and only if the variables are independent and 1 if and only if one is a measurable function of the other, and (c) has a simple asymptotic theory under the hypothesis of independence, like the Pearson correlation. See Section [] for more on this.

The advantages of this new measure motivated us to define a new autocorrelation function using the Chatterjee's correlation coefficent instead of Pearson's. In order to use this in MCMC theory, two things are needed,

1. Some properties that are followed by the classical acf should also hold true for our version for Markov chains (See Section []).

2. As we don't have the luxury of i.i.d. draws, for which the consistency of the estimator of Chatterjee's correlation holds (See []), we need to prove it for the case of MCMC samples we have (See Section []).

We were able to prove three new results related to the Chatterjee's autocorrelation that also hold true for the classical acf. We also believe that

the estimator for Chatterjee's correlation coefficent is consistent even when we're using correlated draws from a stationary Markov chain. We were able to change some parts of the proof of consistency presented in [og paper], and believe that the parts left can also be done and are left as future work.

2. Preliminaries

Definition 2.1. (Time-homogeneous markov chain). The $\mathbb{R}$-valued sequence of random variables $X_1, X_2, \ldots$ is a time-homogeneous Markov chain if for all $A \in \mathcal{B}(\mathbb{R})$ and for all $n \in \mathbb{N}$

$$\Pr(X_{n+1} \in A | X_1, X_2, \ldots, X_n) = \Pr(X_{n+1} \in A | X_n)$$

Definition 2.2. (Markov Transition Kernel). A Markov transition kernel is a map $P : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \to [0, 1]$ such that

1. for all $A \in \mathcal{B}(\mathbb{R}), P(.A)$ is a measurable on $\mathbb{R}$.
2. for all $x \in \mathbb{R}, P(x, .)$ is a probability measure on $\mathcal{B}(\mathbb{R})$.

Definition 2.3. (Stationarity): A discrete-time Markov chain $X_1, X_2, \ldots$ is stationary if the distribution of X_n does not depend on n.

Definition 2.4. (Time Reversibility).

Definition 2.5. (Total Variation Norm).

Definition 2.6. (Ergodicity).

Definition 2.7. (Chapman-Kolmogorov Equation).

Definition 2.8. (Glivenko-Cantelli Theorem).

Theorem 2.9. (Lebesgue's Dominated Convergence Theorem).

3. Problems with Pearson correlation coefficent

Definition 3.1. Pearson correlation coefficient measures the linear correlation of two sets of data. Given a pair of random variables (X, Y), pearson correlation ρ is defined as

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}\,[X]} \cdot \sqrt{\mathrm{Var}\,[Y]}}$$

Pearson's correlation coefficent is very powerful in detecting monotone relations and has a well developed asymptotic theory. The autocorrelation function that we look at in MCMC is also defined using this.

There are two most common problems with this coefficent.

1. Firstly, we would like that the correlation would be close to its maximum value if and only if one variable looks like a noiseless function of the other variable. This is not the case for the Pearson's Coefficient as it is close to ± 1 iff one variable is a noiseless *linear* function.

2. Second, we would like the correlation to be close to its minimum value if and only if both the variables are independent of each other. In the case of the Pearson's correlation, it is zero when the variables are independent but the converse is not always true.

4. Chatterjee's autocorrelation function

Sourav Chatterjee proposed a new correlation coefficient in [add reference]. 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.

Definition 4.1. 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.

The following theorem shows that ξ_n is a consistent estimator of some measure of dependence between the variables X and Y.

Theorem 4.2. 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.

Proof of this theorem is presented in Chatterjee's original paper [og paper ref].

Our aim is to create a Markov chain version of this coefficent, i.e. extend this definition to create an analogous of the autocorrelation function defined using the Pearson correlation. Below are some theorems that'll help us in calculating the Chatterjee's autocorrelation function.

Let $X_1, X_2, \ldots$ be a stationary, time homogeneous Markov chain with stationary distribution π.

Theorem 4.3. Here we present a proof of a well known result about the Pearson Correlation Coefficient.

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

Proof. We know that

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

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

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

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

□

The above result is used in estimating the acf using a single Markov chain. The next theorem is the analogous version of the former for the Chatterjee's correlation.

Theorem 4.4. $\xi(X_n, X_{n+k})$ is independent of n, where n and k are in $\mathbb{N}$.

Proof.

$$\xi_{(X_n, X_{n+k})} = \frac{\int \text{Var}\left[\mathbb{E}[1_{\{X_{n+k} \geq t\}} | X_n = x]\right] d\pi(t)}{\int \text{Var}\left[1_{\{X_{n+k} \geq t\}}\right] d\pi(t)} \tag{4.1}$$

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

$$\mathbb{E}[1_{\{X_{n+k}\geq t\}}|X_n = x] = \Pr(X_{n+k} \geq t|X_n = x)$$

and by time-homogeneity of our Markov chain

$$\Pr(X_{n+k} \geq t|X_n = x) = \int_t^\infty P^k(x, dy)$$

which is independent of n. And hence,

$$\int \operatorname{Var}\left[\int_t^\infty P^k(x, dy)\right] d\pi(u) \tag{4.2}$$

is also independent of n.

Now for the denominator, we know by stationarity of our Markov chain that $X_n \sim \pi$, so for any function f, $f(X_n) \sim \pi'$ for some distribution π', 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

$$\int \operatorname{Var}\left[f_t(X_{n+k})\right] d\pi(t)$$

where,

$$\operatorname{Var}\left[f_t(X_{n+k})\right] = \operatorname{Var}\left[f_t(X_1)\right].$$

Therefore,

$$\int \operatorname{Var}\left[f_t(X_{n+k})\right] d\pi(t)$$

is independent of both n and k.

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

In general, ξ is not a symmetric in X and 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. But for our purposes, it'd be great to have symmmetricity as that is required in estimating the variance in the Markov chain version of the Central Limit Theorem.

Theorem 4.5. $\xi(X_n, X_{n+k}) = \xi(X_{n+k}, X_n)$ for time reversal Markov chains for any $n, k \in \mathbb{N}$.

Proof. By [last theorem ref], we know that the denominator of $\xi(X_n, X_{n+k})$ is independent of both n and k. So we only have to prove that the numerator is symmetric.
We have to show that

$$\int \text{Var}\left[\Pr(X_{n+k} \geq t|X_n)\right] d\pi(t) = \int \text{Var}\left[\Pr(X_n \geq t|X_{n+k})\right] d\pi(t).$$

Lemma 1. *For a time reversible Markov chain, X_n and X_{n+k} are exchangable, i.e.*

$$f_{(X_n, X_{n+k})}(x, y) = f_{(X_{n+k}, X_n)}(x, y) \ \forall (x, y) \in \mathbb{R}^2.$$

Proof. It is enough to show that for any two $A, B \in \mathcal{B}(\mathbb{R})$

$$\Pr(X_n \in A, X_{n+k} \in B) = \Pr(X_{n+k} \in A, X_n \in B)$$

which is same as

$$\int_A \pi(dx) P^k(x, B) = \int_B \pi(dy) P^k(y, A)$$

$$\iff \int_A \int_B \pi(dx) P^k(x, dy) = \int_B \int_A \pi(dy) P^k(y, dx).$$

To prove the above statement, it is enough to show that for any $x \in A$ and $y \in B$,

$$\pi(dx) P^k(x, dy) = \pi(dy) P^k(y, dx).$$

We proceed by strong induction on k. For $k = 1$, it is true by definition of reversibility of Markov chains.

Assume that it is true for all $1 \leq m < k$. We want to prove it for k. By the Chapman-Kolmogorov equation, we have

$$\begin{aligned}\pi(dx)P^k(x,dy) &= \pi(dx)\int_{\mathcal{X}} P^m(x,dz)\cdot P^{k-m}(z,dy)\\ &= \int_{\mathcal{X}} \pi(dx)P^m(x,dz)P^{k-m}(z,dy)\end{aligned}$$

by the inductive hypothesis, we get

$$\begin{aligned}&= \int_{\mathcal{X}} \pi(dz)P^m(z,dx)P^{k-m}(z,dy)\\ &= \int_{\mathcal{X}} P^m(z,dx)\pi(dz)P^{k-m}(z,dy)\\ &= \int_{\mathcal{X}} P^m(z,dx)\pi(dy)P^{k-m}(y,dz)\\ &= \pi(dy)\int_{\mathcal{X}} P^{k-m}(y,dz)P^m(z,dx)\end{aligned}$$

again by the Chapman-Kolmogorov equation, we get that

$$= \pi(dy)\cdot P^k(y,dx).$$

□

By this **Lemma 1**, it is clear that

$$\Pr(X_{n+k} \geq t|X_n) = \Pr(X_n \geq t|X_{n+k})\ \forall t \in \mathbb{R}$$

which implies the result above. □

The next theorem states that as the time difference approaches ∞, the correlation approaches 0.

Theorem 4.6. $\lim_{n\to\infty} \xi(X_1, X_n) = 0$ for an Ergodic Markov chain

Proof. We have

$$\xi(X_1, X_n) = \frac{\int \text{Var}\left[\mathbb{E}[1_{\{X_n \geq t\}}|X_1 = x]\right] d\pi(t)}{\int \text{Var}\left[1_{\{X_n \geq t\}}\right] d\pi(t)}.$$

The denominator is independent of n as proven in [last to last theorem ref], so we only need to show that the numerator goes to 0 as $n \to \infty$.

Lemma 2.

$$\lim_{n\to\infty} \int Var\left[\mathbb{E}[1_{\{X_n \geq t\}}|X_1 = x]\right] d\pi(t) = \int \lim_{n\to\infty} Var\left[\mathbb{E}[1_{\{X_n \geq t\}}|X_1 = x]\right] d\pi(t)$$

Proof. Define $f_n(t) := \text{Var}\left[\Pr(X_n \geq t|X_1 = x)\right] \cdot \pi(t)$.
Assume for the time being that f_n is measurable, $\int_{-\infty}^{\infty} f_n < \infty$ and f_n is continuous.
Now, as f_n is a product of two bounded functions, it is also bounded. Set

$$C := \sup_{n\in\mathbb{N}}(\sup_{t\in\mathbb{R}}(\text{Var}\left[\Pr(X_n \geq t|X_1 = x)\right]))$$

then

$$\int_{-\infty}^{\infty} f_n(t)dt \leq \int_{-\infty}^{\infty} C\pi(t)dt = C < \infty$$

As f_n is dominated by g (where $g(t) := C \cdot \pi(t) \forall t \in \mathbb{R}$),
by Lebesgue's Dominated Convergence Theorem,

$$\lim_{n\to\infty} \int f_n(t)dt = \int (lim_{n\to\infty} f_n(t))dt.$$

□

Lemma 3.

$$\lim_{n\to\infty} Var\left[\mathbb{E}[1_{\{X_n \geq t\}}|X_1 = x]\right] = Var\left[\lim_{n\to\infty} \mathbb{E}[1_{\{X_n \geq t\}}|X_1 = x]\right]$$

Proof. We can write

$$\begin{aligned}\mathrm{Var}\left[\mathbb{E}[1_{\{X_n\geq t\}}|X_1=x]\right] &= \mathbb{E}[\mathbb{E}[1_{\{X_n\geq t\}}|X_1=x]^2] - \mathbb{E}[\mathbb{E}[1_{\{X_n\geq t\}}|X_1=x]]^2 \\ &= \mathbb{E}[\Pr(X_n\geq t|X_1=x)^2] - \mathbb{E}[\Pr(X_n\geq t|X_1=x)]^2.\end{aligned}$$

Assuming that both $\lim_n \mathbb{E}[\Pr(X_n\geq t|X_1=x)^2]$ and $\lim_n \mathbb{E}[\Pr(X_n\geq t|X_1=x)]^2$ exist,

$$\begin{aligned}\lim_{n\to\infty}\mathrm{Var}\left[\mathbb{E}[1_{\{X_n\geq t\}}|X_1=x]\right] &= \lim_{n\to\infty}\mathbb{E}[\Pr(X_n\geq t|X_1=x)^2] \\ &\quad - \lim_{n\to\infty}\mathbb{E}[\Pr(X_n\geq t|X_1=x)]^2 \\ &= \lim_{n\to\infty}\mathbb{E}[\Pr(X_n\geq t|X_1=x)^2] \\ &\quad - (\lim_{n\to\infty}\mathbb{E}[\Pr(X_n\geq t|X_1=x)])^2.\end{aligned}$$

For any $n\in\mathbb{N}$, we can write

$$\mathbb{E}[\Pr(X_n\geq t|X_1=x)^n] = \int_{-\infty}^{\infty}\Pr(X_n\geq t|X_1=x)^n\cdot\pi(t)dt.$$

Lemma 4.

$$\lim_{n\to\infty}\int\Pr(X_n\geq t|X_1=x)^n\cdot\pi(t)dt = \int\lim_{n\to\infty}\Pr(X_n\geq t|X_1=x)^n\cdot\pi(t)dt.$$

Proof. Define $f_n(t) := \Pr(X_n\geq t|X_1=x)^n\cdot\pi(t)$.
Assume for the time being that f_n is measurable, $\int_{-\infty}^{\infty}f_n<\infty$ and f_n is continuous.
Now, as f_n is a product of two bounded functions, it is also bounded.
Now,

$$\int_{-\infty}^{\infty}f_n(t)dt\leq\int_{-\infty}^{\infty}\pi(t)dt=1<\infty.$$

As f_n is dominated by π,
by Lebesgue's Dominated Convergence Theorem,

$$\lim_{n\to\infty}\int f_n(t)dt = \int(\lim_{n\to\infty}f_n(t))dt.$$

□

Using the **Lemma 4** for $n = 1$ and 2, we can take limit in both the terms inside, i.e.

$$\begin{aligned}\lim_{n\to\infty} \text{Var}\left[\mathbb{E}[1_{\{X_n\geq t\}}|X_1 = x]\right] &= \lim_{n\to\infty} \mathbb{E}[\Pr(X_n \geq t|X_1 = x)^2] \\ &- (\lim_{n\to\infty} \mathbb{E}[\Pr(X_n \geq t|X_1 = x)])^2 \\ &= \mathbb{E}[\lim_{n\to\infty} \Pr(X_n \geq t|X_1 = x)^2] \\ &- (\mathbb{E}[\lim_{n\to\infty} \Pr(X_n \geq t|X_1 = x)])^2 \\ &= \text{Var}\left[\lim_{n\to\infty} \mathbb{E}[1_{\{X_n\geq t\}}|X_1 = x]\right]\end{aligned}$$

□

Now, by (2.4), we know that

$$\mathbb{E}[1_{\{X_n\geq t\}}|X_1 = x] = \int_t^\infty P^{n-1}(x, dy)$$

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

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

This implies

$$\begin{aligned}\lim_{n\to\infty} \int \text{Var}\left[\mathbb{E}[1_{\{X_n\geq t\}}|X_1 = x]\right] d\pi(t) &= \lim_{n\to\infty} \int \text{Var}\left[\int_t^\infty P^{n-1}(x, dy)\right] d\pi(t) \\ &= \int \text{Var}\left[\int_t^\infty \lim_{n\to\infty} P^{n-1}(x, dy)\right] d\pi(t) \\ &= \int \text{Var}\left[\int_t^\infty F(dy)\right] d\pi(t) \\ &= \int \text{Var}\left[1 - F(t)\right] d\pi(t) \\ &= \int 0 \cdot d\pi(t) \\ &= 0\end{aligned}$$

under the Total Variation Norm. □

5. Sketch of proof of [Theorem 4.2]

Chatterjee presented the complete proof of [theorem 4.2] in his original paper, where the samples drawn from (X, Y) are i.i.d.
In our case of Markov chains, for the estimation of Chatterjee's correlation, we have correlated but identically distributed draws due to stationarity.
To use the theory in reality, we need to estimate the correlation, and for that we need the estimator to be consistent in our case as well.
Our aim is to prove that the estimator is consistent even in the stationary Markov chain case. We believe that the convergence should happen, but were not able to prove it completely during the UGP timeline.
Below we present [complete the sentence].

6. Some simulation plots and conclusion