APPLICATIONS OF CHATTERJEE'S CORRELATION IN MCMC

Vivek Kumar Singh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
{vksingh}@iitk.ac.in

Dootika Vats
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
{dootika}@iitk.ac.in

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 $X_1, X_2, \dots$ is a 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 3). The autocorrelation function is used for assessing the quality of the Markov chain produced. Pearson's correlation used above is great at detecting monotone relations in data, but doesn't perform well enough when applied to non-linear situations (say $\rho(X, \sin(X))$)

Chatterjee proposed a new measure of dependence in [site paper], which is (a) as simple as the Pearson correlation, (b) 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 4 for more on this.

The advantages of this new measure motivated us to define a new autocorrelation function using Chatterjee's correlation coefficient instead of Pearson's. In order to use this in MCMC theory, two things were needed,
1. Some properties that are followed by the classical ACF should also hold for our version for Markov chains.
2. As we do not have the luxury of i.i.d. draws, for which the consistency of the estimator of Chatterjee's correlation holds (See [paper]), we need to prove it for the case of MCMC samples we have.
We proved three results for Chatterjee's correlation that are analogous to their Pearson counterpart (See Section 5). We also believe that the estimator for Chatterjee's correlation coefficient is consistent even when we are using correlated draws from a stationary Markov chain. Some ideas are presented (See Section 6) inspired by the proof of consistency in [paper] of the i.i.d. estimator on how one should go about it. We were not able to complete it in this project and is left as future work.

2 Markov chain Monte Carlo

A Markov chain is a discrete time stochastic process $X_1, X_2, \dots$ taking values in an arbitrary general state space and having the Markov property, i.e. the conditional distribution of X_{n+1} given the past $X_1, \dots X_n$, depends only on the present state X_n.

A Markov chain can be specified by two things, the intital distribution and the transition probabilities. The initial distribution is the marginal of X_1. By transition probabilities, we mean the conditional distribution of X_{n+1} given X_n. We always assume stationary transition probabilities, i.e. the conditional $X_{n+1}|X_n$ is independent of n.
A Markov chain is said to be stationary if the distribution of X_n is independent of n, and ergodic if it converges to the invariant distribution regardless of the choice of the initial distribution.
A stationary Markov chain is time-reversible w.r.t. the stationary distribution if X_n and X_{n+1} are exchangable.

3 Pearson's Correlation Coefficient

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

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

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

3.1 Problems with Pearson correlation

There are two most common problems with this coefficent.

1. First, 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 as ρ 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, it is zero when the variables are independent but the converse is not always true.

4 Chatterjee's Correlation Coefficient

In order to solve these problems, Chatterjee came up with a new correlation coefficient which overcomes these drawbacks and also has a consistent estimator which is computationally efficient.
Chatterjee's correlation ξ is defined as

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

where X, Y are any sets of random variables.

4.1 A consistent estimator of ξ

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 4.1. *If Y is not almost surely a constant, then as $n \to \infty$, $\xi_n(X, Y)$ converges almost surely to $\xi(X, Y)$, where μ is the pdf of Y.*

This [theorem] proves that ξ_n is a consistent estimator of ξ. Proof is given in [og paper].

4.2 Properties

Some properties of this new measure of dependence are as follows

1. $\xi(X, Y) \in [0, 1]$
2. $\xi(X, Y) = 0$ if and only if X and Y are independent.
3. $\xi(X, Y) = 1$ if and only if atleast one of X and Y is a measurable function of the other.
4. 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 $\max(\xi_n(X, Y), \xi_n(Y, X))$.
5. ξ_n is based on ranks, and for the same reason, it can be computed in $O(n \log n)$.

5 Chatterjee's Autocorrelation Function

The aim of this project is to create a new autocorrelation function (ACF) using the Chatterjee's correlation coefficient, and study it's behaviour when applied to Markov chains.
Let $X_1, X_2, \ldots$ be a stationary, time homogeneous Markov chain with stationary distribution π. We define the new lag-k ACF as follows

$$\gamma'_{k,n} := \xi(X_n, X_{n+k}).$$

Theorem 5.1. *Here we present a proof of a well known result about the Pearson Correlation Coefficient.*

$$\textit{Cov}(X_n, X_{n+k}) \textit{ is independent of } n.$$

Proof. We know that

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

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

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

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

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

As both the terms of $\text{Cov}(X_n, X_{n+k})$ are independent of n, $\text{Cov}(X_n, X_{n+k})$ is independent of n.

□

This theorem is important as it allows us to estimate the ACF without drawing multiple Markov chains. This suggests that the same property should also hold for $\gamma'_{k,n}$, proved in the next theorem.

Theorem 5.2. $\xi(X_n, X_{n+k})$ *is independent of* n*, where* n *and* k *are in* $\mathbb{N}$*.*

Proof.

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

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,

$$\int \text{Var}\left[\int_t^\infty P^k(x, dy)\right] d\pi(u) \tag{2}$$

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

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

where,

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

Therefore,

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

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. □

Because of Theorem 5.2, we can safely denote this new ACF by γ'_k.

Now, if the Central Limit Theorem holds for a Markov chain, the variance is as follows

$$\sigma^2_{\text{clt}} = \sum_{k=-\infty}^{\infty} \gamma_k.$$

As γ_k is symmetric, we get that

$$\sigma^2_{\text{clt}} = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k.$$

Now ξ is not symmetric in general, but for our purposes, it'd be great to have symmmetricity so that we can do a similar step as above. We've proved the symmmetricity of ξ in the case of time reversible Markov chains.

Theorem 5.3. *$\xi(X_n, X_{n+k}) = \xi(X_{n+k}, X_n)$ for time reversal Markov chains for any $n, k \in \mathbb{N}$.*

Proof. By Theorem 5.2, 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

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

Lemma 5.4. *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 Lemma 5.4, 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. □

For an Ergodic Markov chain, we can intuitively argue that the correlation should approach 0 as the time difference increases. In the next theorem, we've proved that this holds true for ξ.

Theorem 5.5. $\lim_{n\to\infty} \xi(X_1, X_n) = 0$ *for an Ergodic Markov chain*

Proof. We have

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

The denominator is independent of n as proven in Theorem 5.2, so we only need to show that the numerator goes to 0 as $n \to \infty$.

Lemma 5.6.

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

Proof. Define $f_n(t) := \text{Var}[\Pr(X_n \geq t|X_1 = x)] \cdot \pi(t)$.
It is easy to see 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

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

then

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

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,

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

□

Lemma 5.7.

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

Proof. We can write

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

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,

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

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 5.8.

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

Proof. Define $f_n(t) := \Pr(X_n\geq t|X_1=x)^n\cdot\pi(t)$.
It is easy to see 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,

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

□

Using Lemma 5.8 for $n=1$ and 2, we can take limit in both the terms inside, i.e.

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

□

Now 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 for every initial distribution λ

$$\lim_{k \to \infty} ||\lambda P^k - F|| = 0.$$

Consequently, for all $x \in \mathbb{R}$,

$$\lim_{k \to \infty} ||P^k(x, \cdot) - F(\cdot)|| = 0.$$

So, define

$$G_{k,x}(t) := \int_{-\infty}^t P^k(x, dy)$$

i.e. the cdf of $P^k(x, \cdot)$. By ergodicity, we get that

$$G_{k,x}(t) \xrightarrow{d} F(t)$$

This implies

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

□

6 Proof sketch of Theorem 4.1 for correlated samples

Chatterjee presented the complete proof of Theorem 4.1 in [paper], where the samples drawn from (X, Y) are i.i.d. In our case of MCMC, for the estimation of Chatterjee's correlation, we have correlated but identically distributed draws due to stationarity. To use this new autocorrelation function, we need to estimate it first, 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 does happen, but were not able to prove it completely in this project, and it is left as future work.

Theorem 6.1. *Let $X_1, X_2, \ldots$ be a stationary time-homogeneous Markov chain with stationary distribution μ. Then $\xi_n(X, Y)$ estimated using the draws from the Markov chain converge to $\xi(X, Y)$ as $n \to \infty$, where X, Y are any two time points in the chain.*

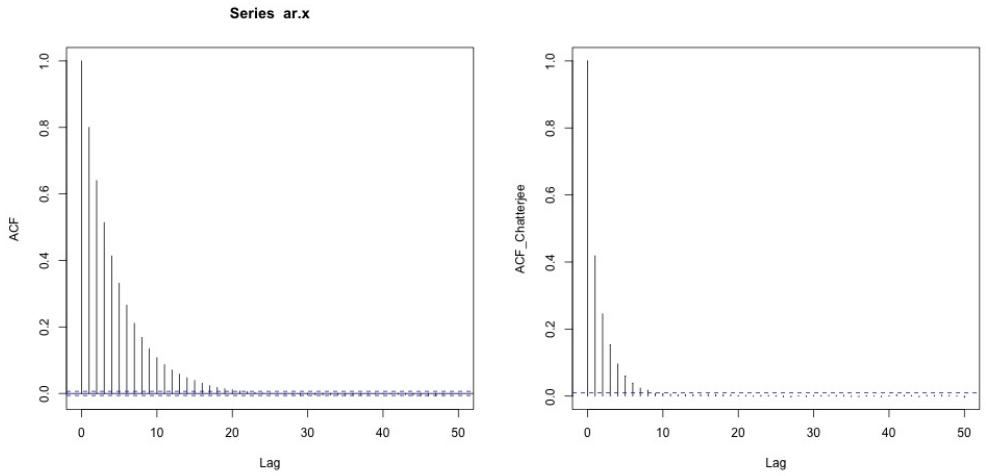
All the Lemmas and Corollaries are taken directly from the proof in the i.i.d. case given in [paper].

Proof. □

7 Simulations

Assuming Theorem 6.1, we've presented below two sets of ACF plots of two different Markov chains. On the left column, we have Pearson's ACF plot, and on the right we have Chatterjee's ACF plot.

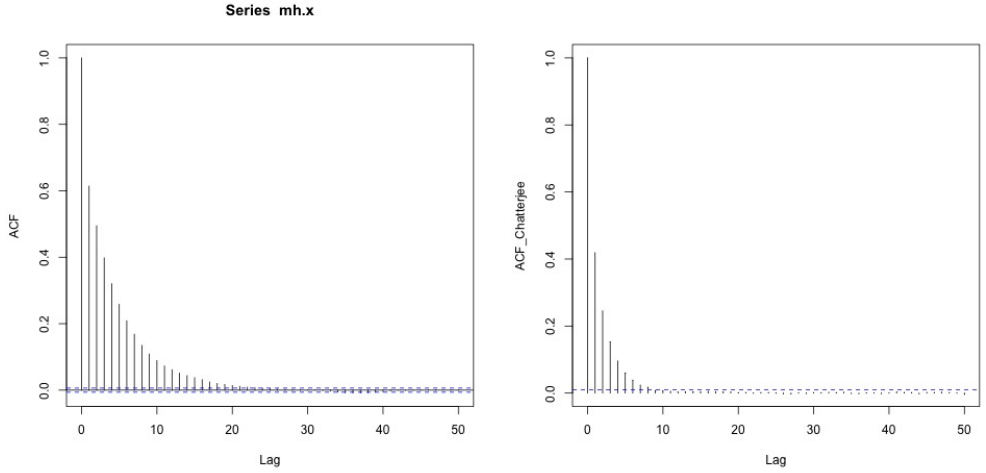
First, we have the AR(1) process with $\rho = 0.8$.



(a) Classical ACF (b) Chatterjee's ACF

Figure 1: AR(1) Process with $\rho = 0.8$

Next, we have the Metropolis-Hastings algorithm with initial distribution as Exp(0.01) and target distribution $\mathcal{N}(0, 1)$



(a) Classical ACF (b) Chatterjee's ACF

Figure 2: Metropolis Hastings algorithm with intial distribution Exp(0.01) and target distribution $\mathcal{N}(0, 1)$

In both the cases, we can see that the rate of convergence of ACF is faster for the Chatterjee's correlation. If in future we can find a relation between variance of Markov chain CLT and Chatterjee's autocorrelation, then this faster convergence rate might make it a better alternative in estimating the variance.

References

8 Appendix