APPLICATIONS OF CHATTERJEE'S CORRELATION IN MCMC

Vivek Kumar Singh
IITK
Indian Institute of Technology Kanpur
{vksingh}@iitk.ac.in

Dootika Vats
IITK
Department of Mathematics and Statistics
{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 $X_1, X_2, \ldots$ 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. 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 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 Chatterjee's correlation coefficient 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 for our version for Markov chains (See Section []).
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 []), we need to prove it for the case of MCMC samples we have (See Section []).
We proved three results for Chatterjee's correlation that are analogous to their Pearson counterpart (See Section []). 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 []) inspired by the proof of consistency in [] 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

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 $O(n \log n)$.