Yumo Peng
yp30751@uga.edu

Estimation of Errors in the Mean in Serial Data with Autocorrelation

Foley, B. L.

Complex Carbohydrate Research Center, University of Georgia

and

Peng, Y.

Department of Statistics, University of Georgia

and

Lazar, N. A.

Department of Statistics, The Pennsylvania State University

November 23, 2020

Abstract

We present here an elaboration on certain existing techniques for estimating errors in means of serial data with autocorrelation. Data from molecular dynamics (MD) simulations are the primary motivation for this study because they present certain unusual statistical challenges. For that reason, most examples herein address MD data, but the techniques are applicable to any data with similar characteristics. In this elaboration, we: modify the earlier procedures to simplify their application; briefly explore theoretical implications of the method; present and discuss example applications; and suggest an algorithm that can be applied both during and after data acquisition. Additionally, we show that the method can be used to interrogate data for the presence of steady-state phenomena, e.g., equilibration and convergence, with respect to a monitored quantity, and suggest a procedure for establishing whether a quantity of data is sufficient for meaningful statistical descriptions.

Keywords: Statistical inefficiency, Blocking, Equilibration, Convergence, Correlation length

Blocking Method for Error Estimation in Molecular Dynamics and its Significance

BACKGROUND

Molecular dynamics is a simulation method used for mimicking the actual physical movement of atoms and molecules. Using a single MD simulation to model the motion of water molecules, we can get a large amount of data recording the instantaneous distance between two hydrogen atoms. To have a preliminary understanding of this huge data set, computing the average tells us the mean distance between the hydrogen atoms. What should we expect for the mean distance from the next simulation? As molecules and atoms keep moving, we will observe different information about the distance between two hydrogen atoms for each simulation, which indicates the mean distance to be a fluctuating quantity as well. Therefore, although knowing the sample mean could tell us the accuracy of the simulated data set by comparing how close the sample mean is to the population mean, making any conclusions based only on the individual sample mean is misleading. It is necessary to understand how the sample mean varies and to investigate a reliable statistic to quantify the precision of the sample means derived from each simulation.

In statistics, variance in means (V_m), measuring the variation of the sample mean around the population mean($\mathrm{V_x}$), is commonly used for quantifying statistical uncertainty (statistical error).

RESEARCH QUESTION and IDEA

Traditionally, for uncorrelated data, it is easy to compute $\mathrm{V_m} = \frac{\mathrm{V_x}}{N}$. However, MD simulation data are time series data with autocorrelation. We can think of them as a type of data that has “memory,” where not every data point provides the same amount of new information, and the information can also be explained by data from previous time points. Therefore, it is necessary to include a new quantity "Statistical inefficiency" (I) to the calculation of $\mathrm{V_m} = \frac{(\mathrm{I} * \mathrm{V_x})}{N}$. Unfortunately, computing I is time-consuming. This difficulty then leads to a key research question important for further investigation: How can we get an appropriate estimate for V_m without considering the value of I?

METHOD

We start with dividing our data (from a single MD simulation) into b blocks. Each block should contain nearly equal numbers of consecutive observations. Then, $V_m(b)$ over the b blocks are calculated. For each value of b, we should get a corresponding $V_m(b)$ and plot $V_m(b)$ versus b. From the plot, there are two criteria to determine the proper estimate of V_m. For simple autocorrelated data without oscillations, we look for the “flat region” where $V_m(b)$ starts leveling off to an asymptotic value, which is the estimate of $\mathrm{V_m}$. For autocorrelated data with oscillations, instead, we look for the “bump region” where the $V_m(b)$ reaches the maximum, and that is the proper estimate of $\mathrm{V_m}$.

IMPACTS

The most optimal method to determine $\mathrm{V_m}$ is to run multiple simulations and calculate the average of the variance from each simulation. However, this is not the best way, especially for MD simulation, which requires a large amount of time and computational power even for a single simulation. Another method is to directly use the formula $\frac{(\mathrm{I}*\mathrm{V}_\mathrm{x})}{N}$, but we indicated the difficulty of knowing the value of I. The blocking method addresses these two concerns. It requires only a single simulation data and could provide appropriate estimates of V_m without any calculation including I, which makes the error estimation process less computationally intensive and more efficient.

MD simulation and error estimation also have significant social impacts. It has been widely applied in the molecular structure simulation and drug design and discovery. Take the Coronavirus as an example: at first, while we might not know anything about the virus, MD simulation could be used to model the molecular structure of it; when the molecular structure of the virus is known, MD simulation could further explore the key sites (biological target) of the molecule, which is helpful for discovering and designing inhibitors to inhibit and control the reproduction of the virus. Among all of the applications of MD simulation, taking the statistical error (V_m) into consideration with the simulated results and obtaining a reliable estimate of the error becomes a necessary issue in the simulation process because poor estimate that underestimates the true statistical error could lead to incorrect conclusion. Therefore, as the application of MD simulation becomes more popular, the blocking method will also play an important role in those applications since it efficiently provides us a proper error estimation and MD simulation is always subject to statistical error.

Research Progress Report

Yumo Peng

Mentors: Dr. Lachele Foley; Professor Nicole Lazar

December 9, 2020

Contents

1 Introduction

The project is meant to develop an approach that could more efficiently solve the statistical challenge of estimating the variance in the mean (VIM) in time series data.

We have already applied the blocking method on many samples of MD simulated data, and it is worth of knowing how well the blocking method will be performed on other types of serial data. In the meantime, we are also interested in applying a new optimization method called "bump-hunting" to the data, so here also includes descriptions, implementation of the method and output comparison with blocking method and the theoretical true value.

2 Works Completed

Pseudo data Generation:

Using R, I generated 81 samples of pseudo data from AR(1) process, with 9 different white noise standard deviation(2 to 9) and 9 different AR coefficients(0.1 to 0.9). The sample size of each sample is 50,000. The corresponding 81 theoretical variance of means were also calculated.

Blocking method:

We start with dividing our data (AR(1) simulating data) into b blocks. Each block should contain nearly equal numbers of consecutive observations. Then, $V_m(b)$ over the b blocks are calculated. For each value of b, we should get a corresponding $V_m(b)$ and plot $V_m(b)$ versus b. From the plot, there are two criteria to determine the proper estimate of V_m. For simple autocorrelated data without oscillations, we look for the "flat region" where $V_m(b)$ starts leveling off to an asymptotic value, which is the estimate of V_m. For autocorrelated data with oscillations, instead, we look for the "bump region" where the $V_m(b)$ reaches the maximum, and that is the proper estimate of V_m.

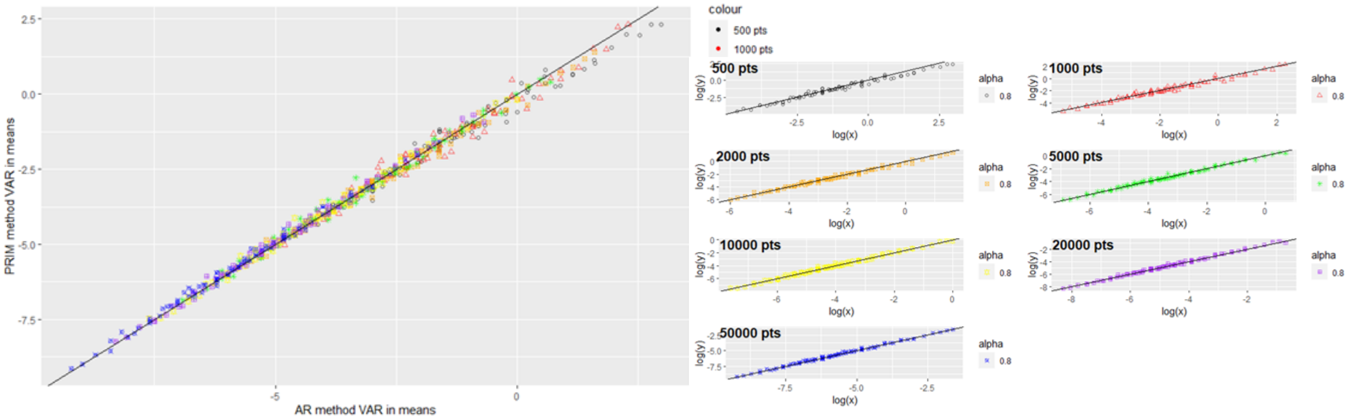
Even though blocking method is a more efficient method of estimating the VIM compared to the theoretical computation method, the way we get the proper estimate of VIM is merely through looking at the plot, which makes the process of observing and recording the estimate of VIM to be quite subjective. In order to obtain better estimates in a more computational way, we integrated using bump-hunting method, which was introduced by Friedman & Fisher(1999).

Bump-hunting method:

General Idea of Bump-hunting: To seek a set of sub-regions of the input variable space within which the value of the output variable is considerably larger (or smaller) than its average value over the entire input domain. We start with a box that covers all of the data, and at each step, one small sub-region(determined by α) will be removed from the current box. The box remained is the one that will produce the largest response mean value within the next sub-region after removal. For each iteration, the current box will then be updated, and the peeling procedure will repeat until we get a newer smaller sub-region box. The "peeling" process should continue until we have β fraction points remain.

1) Run bump-hunting method on all simulation data; Make comparison plots and visually see how well the estimation is:

I made a scatterplot(in log-scale) with x be the theoretical variance in means and y be the suggested variance in means by the bump-hunting method. There are 7 different scatterplots overlaid together, with each one corresponding to different number of points are used for bump-hunting method. Nearly all of the points are fall along with the straight-line y = x, indicating that the estimation performs well.

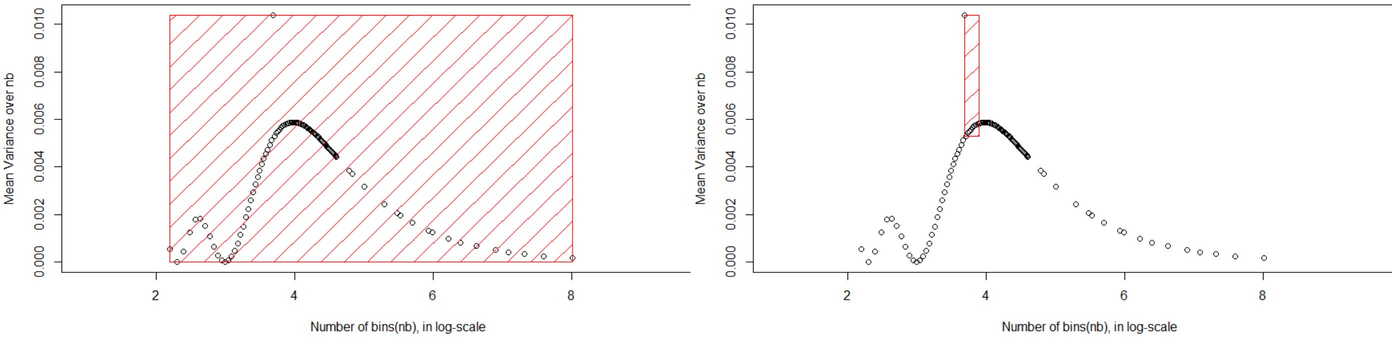


(a) Comparison Plot using 81 samples of data

2) Run Bump-hunting on other types of serial data (sin-waves data); Update Bump-hunting method for handling outliers:

We were also interested in knowing how the bump-hunting method will perform on other types of serial data, so I also applied the method on Sine-Wave data and see if the method does catch the bump region successfully.

Based on one of the example plot shown below, the bump-hunting seeks the bump region successfully, as indicated by the red enclosed region. During the process of testing other samples of data, we do notice that for some data set, there might be potential outliers with high value around the bump region, which could possibly affect catching the real bump region. Thus, we updated the current algorithm by including steps of using IQR(Interquartile Range) condition to check the existence of outliers, and if exists, they will be removed before starting the bump-hunting.



3) Write tutorials describing the method and simulation process:

Since our proposed method is also designed for non-specialists that are not in Statistics or Chemistry field, I created comprehensive and step-by-step tutorials guiding the user to generate their own simulation AR(1) data and how to implement the bump-hunting algorithm in R. Users could choose their own peeling parameters based on their own needs. Moreover, users could either see the whole data peeling process and print out the intermediate bump-regions or directly print out the eventual optimal bump region and use it to compute the estimated VIM.

3 Plans for Next Semester

1) Summarize every finished or in-process phased tasks, starting from the beginning up to now

2) Prepare for making the poster for the CURO conference presentation

3) Figure out how to do the calculation and derivative using Maple

4) Reference for the pre-published research paper