Robust estimations from distribution structures: III. Non-asymptotic

immediate

This manuscript was compiled on November 25, 2023

Due to the hardness of order statistics, finite sample bias of robust statistics is generally unsolvable. Monte Carlo method can provide approximate solutions, but the convergence rate is very low, so the computational cost to achieve desired accuracy is unaffordable for ordinary users. Here, with a comparametric approach, the distribution structure of order statistics is decomposed. By obtaining a set of quasi-random variables simultaneously consistent for two or more different estimators, the finite sample bias of other related estimators can be approximated with much less computational costs. This article provides a different and prospective approach to integrate two or more parametric assumptions.

finite sample bias | order statistics | variance reduction | Monte Carlo study

The simplest robust estimator, median, has a very complex finite sample behavior. If n is odd, $E\left[median_n\right] = \int_{-\infty}^{\infty} \left(\frac{n+1}{2}\right) \left(_{\frac{n}{2}-\frac{1}{2}}^{\ n}\right) F\left(x\right)^{\frac{n}{2}-\frac{1}{2}} \left[1-F(x)\right]^{\frac{n}{2}-\frac{1}{2}} f(x) x dx$ (**?**), where $F(x)$ and $f(x)$ are the cdf and pdf of the assumed distribution. For the exponential distribution, the above equation is analytically solvable, i.e., $E\left[median_n\right] = \frac{2^{-n-1}(n+1)\left(_{\frac{n-1}{2}}^{\ n}\right)\left(H_n - H_{\frac{n-1}{2}}\right)\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi}}{\lambda\Gamma\left(\frac{n}{2}+1\right)}$, where H_n is the nth Harmonic number, Γ is the gamma function, λ is the scale parameter of the exponential distribution. However, for distributions having more complicated pdf, such equations are generally unsolvable. Even a numerical solution requires exponential time complexity (**? ?**). So, Monte Carlo simulation is currently the only practical choice to estimate finite sample bias. However, the computational cost is too high to be processed in a typical PC. Usually, for robust scale estimators, about 1 million pseudorandom samples for each value of n is needed to ensure three decimal accuracy (**? ?**). In addition to computational challenges, there is an inherent difficulty of dealing with randomness. The theory of probability provides the framework to model and understand random phenomena, but the practical implementation of these models can be challenging. The quality of randomness can affect the validity of the simulation results. Thus, it is not surprised that currently the only popular exact finite sample bias correction is the factor for unbiased standard deviation for the Gaussian distribution, which can be deduced theoretically (**?**). The purpose of this article is to show that the uniform random variables can be decomposed using a few quasi-random variables with high accuracy and the computational cost of finite sample bias estimation from Monte Carlo study can be dramatically improved by obtaining a set of quasi-random variables simultaneously consistent for two or more different estimators.

Any continuous distribution can be linked to the uniform distribution on the interval [0, 1] through its quantile function. This fundamental concept in Monte Carlo study implies that understanding the structure of uniform random variables can be leveraged to understand the structure of any other continuous random variable through the quantile transform. Consider if we use a series of quasi-random variables to approximate the structure of uniform random variables, then the most nature choice is the arithmetic sequence, i.e., $\{x_i\}_{i=1}^n = \left\{\frac{i}{n+1}\right\}_{i=1}^n$. However, the bias of the arithmetic central moments estimated from the Gaussian distribution obtained from arithmetic sequence is very different from that obtained from the pseudo-random variables (Figure 1). The arithmetic sequence lacks the variability of true random samples. A random sample from a uniform distribution would have central moments that could be estimated unbiasedly with sufficient sample size, but an arithmetic sequence would not capture this variability. To better replicate the features of uniform random variables, we introduced beta distributions with a variety of parameters and weights, resulting in distributions that are U-shaped, n-shaped, left-skewed, and right-skewed. By using constraint optimization to assign weights to these distributions, this approach enhanced the approximation of uniform random variables, as demonstrated in Figure 1. The findings suggest that arithmetic sequences account for approximately 60% of the properties of uniform random variables, while beta distributions with different shapes each contribute about 10%. Consequently, with about 90% precision, uniform random variables can be decomposed using just five quasi-random variables, an approach analogous to the Fourier transformation.

To further increase precision, we adopted a stochastic method: pseudo-randomly generating twelve sequences and evaluating their efficacy in approximating uniform random variables by estimating the biases in central moment estimations for Gaussian and exponential distributions. Sequences meeting the predetermined accuracy threshold were retained, as detailed in Algorithm 1; those that did not were discarded in favor of a new set. Upon identifying twenty sequence sets that accurately approximate uniform random variables, these

Significance Statement

In statistics, most current theories focus on asymptotic analysis due to its tractability and simplicity. Non-asymptotic statistics are crucial when dealing with small or moderate sample sizes, which is often the case in practice. Monte Carlo studies are a powerful tool for dealing with non-asymptotic behavior where analytical results are difficult or impossible to obtain. However, these studies can be computationally expensive, especially if high precision is required, or if the statistical model requires significant computational time. Here, we propose calibrated Monte Carlo study, which aims to approximate the randomness structures with a small set of quasi-random variables. This approach sheds light on understanding the general structure of randomness.

sets were applied to assess the finite sample biases in other moment estimators, such as the median, the Hodges-Lehmann estimator, and the standard deviation. The outcomes indicate that using merely twenty sets of sequences, which can be executed on a standard PC in a negligible amount of time, achieves a precision of approximately 0.005. By contrast, attaining the same level of precision using classic Monte Carlo methods would require roughly 0.1 million pseudo-random samples.

Methods

Data and Software Availability

All data are included in the brief report and SI Dataset S1. All codes have been deposited in GitHub.