Semiparametric robust mean estimations based on the orderliness of quantile averages

Tuban Lee

This manuscript was compiled on May 24, 2023

As one of the most fundamental problem in statistics, robust location estimation has many prominent solutions, such as the symmetric trimmed mean, symmetric Winsorized mean, Hodges–Lehmann estimator, Huber M-estimator, and median of means. Recent studies suggest that their maximum biases concerning the mean can be quite different in asymmetric distributions, but the underlying mechanisms and average performance remain largely unclear. This study establishes several forms of orderliness among quantile averages, similar to the mean-median-mode inequality, within a wide range of semiparametric distributions. From this, a sequence of advanced robust mean estimators emerges, which also explains why the Winsorized mean and median of means typically have smaller biases compared to the trimmed mean. Building on the U-orderliness, the superiority of the median Hodges–Lehmann mean is discussed.

semiparametric | mean-median-mode inequality | asymptotic | unimodal | Hodges–Lehmann estimator

In 1823, Gauss (1) proved that for any unimodal distribution with a finite second moment, $|m-\mu| \leq \sqrt{\frac{3}{4}}\omega$, where μ is the population mean, m is the population median, and ω is the root mean square deviation from the mode, M. This pioneering work revealed that despite potential bias with respect to the mean in robust estimates, the deviation remains bounded in units of a scale parameter under a semi-parametric assumption. Bernard, Kazzi, and Vanduffel (2020) (2) further derived asymptotic bias bounds of any quantile for unimodal distributions with finite second moments by reducing this optimization problem to a parametric one, which can be solved analytically. They showed that the population median, m, has the smallest maximum distance to the population mean, μ, among all symmetric quantile averages (SQA_ϵ). Daniell, in 1920, (3) analyzed a class of estimators, linear combinations of order statistics, and identified that ϵ-symmetric trimmed mean (STM_ϵ) belongs to this class. Another popular choice, the ϵ-symmetric Winsorized mean (SWM_ϵ), named after Winsor and introduced by Tukey (4) and Dixon (5) in 1960, is also an L-estimator. Bieniek (2016) derived exact bias upper bounds of the Winsorized mean based on Danielak and Rychlik's work (2003) on the trimmed mean for any distribution with a finite second moment and confirmed that the former is smaller than the latter (6, 7). In 1963, Hodges and Lehmann (8) proposed a class of nonparametric location estimators based on rank tests and, from the Wilcoxon signed-rank statistic (9), deduced the median of pairwise means as a robust location estimator for a symmetric population. Both L-statistics and R-statistics achieve robustness essentially by removing a certain proportion of extreme values. In 1964, Huber (10) generalized maximum likelihood estimation to the minimization of the sum of a specific loss function, which measures the residuals between the data points and the model's parameters. Some L-estimators are also M-estimators, e.g., the sample mean is an M-estimator with a squared error loss function, the sample median is an M-estimator with an absolute error loss function (10). The Huber M-estimator is obtained by applying the Huber loss function that combines elements of both squared error and absolute error to achieve robustness against gross errors and high efficiency for contaminated Gaussian distributions (10). Sun, Zhou, and Fan (2020) examined the concentration bounds of Huber M-estimator (11). Mathieu (2022) (12) further derived the concentration bounds of M-estimators and demonstrated that, by selecting the tuning parameter which depends on the variance, Huber M-estimator can also be a sub-Gaussian estimator. The concept of median of means ($\mathrm{MoM}_{k,b=\frac{n}{k}}$, k is the number of size in each block, b is the number of blocks) was implicitly introduced several times in Nemirovsky and Yudin (1983) (13), Jerrum, Valiant, and Vazirani (1986), (14) and Alon, Matias and Szegedy (1996) (15)'s works. Given its good performance even for distributions with infinite second moments, MoM has received increasing attention over the past decade (16–18). Devroye, Lerasle, Lugosi, and Oliveira (2016) showed that MoM nears the optimum of sub-Gaussian mean estimation with regards to concentration bounds when the distribution has a heavy tail (17). For a comparison of concentration bounds of trimmed mean, Huber M-estimator, median of means and other relavent estimators, readers are directed to Gobet, Lerasle, and Métivier's paper (2022) (19). Laforgue, Clemencon, and Bertail (2019) proposed the median of randomized means ($\mathrm{MoRM}_{k,b}$) (18), wherein, rather than partitioning, an arbitrary number, b, of blocks are built independently from the sample, and showed that MoRM has a better non-asymptotic sub-Gaussian property compared to MoM. In fact, asymptotically, the Hodges-Lehmann (H-L) estimator is equivalent to $\mathrm{MoM}_{k=2,b=\frac{n}{k}}$ and $\mathrm{MoRM}_{k=2,b}$, and they can be seen as the pairwise mean distribution is approximated by the sampling without replacement and bootstrap, respectively. For the asymptotic validity, readers are referred to the foundational works of Efron (1979) (20), Bickel and

Significance Statement

In 1964, van Zwet introduced the convex transformation order for comparing the skewness of two distributions. This paradigm shift played a fundamental role in defining robust measures of distributions, from spread to kurtosis. Here, instead of examining the stochastic ordering between two distributions, the orderliness of quantile averages within a distribution is investigated. By classifying distributions through the signs of derivatives, a series of sophisticated robust mean estimators are deduced. Nearly all common nonparametric robust location estimators are found to be special cases thereof.

T.L. designed research, performed research, analyzed data, and wrote the paper.

The author declares no competing interest.

[1] To whom correspondence should be addressed. E-mail: tl@biomathematics.org

Freedman (1981, 1984) (21, 22), and Helmers, Janssen, and Veraverbeke (1990) (23).

Data Availability. Data for Figure **??** are given in SI Dataset S1. All codes have been deposited in GitHub.

ACKNOWLEDGMENTS. I sincerely acknowledge the insightful comments from the editor which considerably elevated the lucidity and merit of this paper.

1. CF Gauss, *Theoria combinationis observationum erroribus minimis obnoxiae*. (Henricus Dieterich), (1823).
2. C Bernard, R Kazzi, S Vanduffel, Range value-at-risk bounds for unimodal distributions under partial information. *Insur. Math. Econ.* **94**, 9–24 (2020).
3. P Daniell, Observations weighted according to order. *Am. J. Math.* **42**, 222–236 (1920).
4. JW Tukey, A survey of sampling from contaminated distributions in *Contributions to probability and statistics*. (Stanford University Press), pp. 448–485 (1960).
5. WJ Dixon, Simplified Estimation from Censored Normal Samples. *The Annals Math. Stat.* **31**, 385 – 391 (1960).
6. K Danielak, T Rychlik, Theory & methods: Exact bounds for the bias of trimmed means. *Aust. & New Zealand J. Stat.* **45**, 83–96 (2003).
7. M Bieniek, Comparison of the bias of trimmed and winsorized means. *Commun. Stat. Methods* **45**, 6641–6650 (2016).
8. J Hodges Jr, E Lehmann, Estimates of location based on rank tests. *The Annals Math. Stat.* **34**, 598–611 (1963).
9. F Wilcoxon, Individual comparisons by ranking methods. *Biom. Bull.* **1**, 80–83 (1945).
10. PJ Huber, Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
11. Q Sun, WX Zhou, J Fan, Adaptive huber regression. *J. Am. Stat. Assoc.* **115**, 254–265 (2020).
12. T Mathieu, Concentration study of m-estimators using the influence function. *Electron. J. Stat.* **16**, 3695–3750 (2022).
13. AS Nemirovskij, DB Yudin, *Problem complexity and method efficiency in optimization*. (Wiley-Interscience), (1983).
14. MR Jerrum, LG Valiant, VV Vazirani, Random generation of combinatorial structures from a uniform distribution. *Theor. computer science* **43**, 169–188 (1986).
15. N Alon, Y Matias, M Szegedy, The space complexity of approximating the frequency moments in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. pp. 20–29 (1996).
16. PL Bühlmann, Bagging, subagging and bragging for improving some prediction algorithms in *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule (ETH)*. (Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich), Vol. 113, (2003).
17. L Devroye, M Lerasle, G Lugosi, RI Oliveira, Sub-gaussian mean estimators. *The Annals Stat.* **44**, 2695–2725 (2016).
18. P Laforgue, S Clémençon, P Bertail, On medians of (randomized) pairwise means in *International Conference on Machine Learning*. (PMLR), pp. 1272–1281 (2019).
19. E Gobet, M Lerasle, D Métivier, Mean estimation for Randomized Quasi Monte Carlo method. working paper or preprint (2022).
20. B Efron, Bootstrap methods: Another look at the jackknife. *The Annals Stat.* **7**, 1–26 (1979).
21. PJ Bickel, DA Freedman, Some asymptotic theory for the bootstrap. *The annals statistics* **9**, 1196–1217 (1981).
22. PJ Bickel, DA Freedman, Asymptotic normality and the bootstrap in stratified sampling. *The annals statistics* **12**, 470–482 (1984).
23. R Helmers, P Janssen, N Veraverbeke, *Bootstrapping U-quantiles.* (CWI. Department of Operations Research, Statistics, and System Theory [BS]), (1990).