*Write a scientific introduction to the parallel computing procedure you have chosen, explain what the algorithm does and why it works.*

As computing power has dramatically increased in recent decades, the use of bootstrapping as a fundamental data analysis tool has also surged.[[1]](#endnote-1) Bootstrapping is a general statistical technique for estimating unknown quantities utilizing random resampling with replacement for sample data.[[2]](#endnote-2) This method is often used to approximate standard errors, confidence intervals, and probability values for test statistics for a given distribution of data.[[3]](#endnote-3) The term “bootstrapping” was coined by Bradley Efron in 1979 as a reference to the adage of “pulling oneself up by one’s own bootstraps” since the technique approximates statistical figures utilizing the sample data themselves instead of any external assumptions.[[4]](#endnote-4)

The key approach behind bootstrapping is the random sampling of a given number of values from sample data with replacement.[[5]](#endnote-5) Sampling with replacement means that selected values are not removed from the distribution, which allows certain values to be selected multiple times while other values may not be selected at all.[[6]](#endnote-6) This process maintains the data structure while reshuffling the values to calculate sample statistics and ultimately estimate the population statistics.[[7]](#endnote-7) Since biases can be present in this process, statisticians have attempted to provide more accurate results with procedural variants, including parametric bootstrapping, nonparametric bootstrapping, and Bayesian bootstrapping.[[8]](#endnote-8)

Although the resampling of sample data may not intuitively seem to provide new insights, bootstrapping works by avoiding assumptions about the population distribution and utilizing the only information available, the sample data, to create a distribution that may closely resemble that of the population. [[9]](#endnote-9) Further theoretical justifications for bootstrapping have been described in by Peter Hall in “On Bootstrap Confidence Intervals in Nonparametic Regression” (1992)[[10]](#endnote-10), Dimitris N. Politis, Joseph P. Romano, and Michael Wolf in “Subsampling” (1999)[[11]](#endnote-11), and S. N. Lahiri in “Resampling Methods for Dependent Data” (2003)[[12]](#endnote-12).

Bootstrapping has several methodological advantages. The process enables statistical inferences to be drawn from small samples when additional information is not available, which is often the case due to the time and cost of gathering more data.[[13]](#endnote-13) The approach also does not make assumptions about the data distributions and can consequently be helpful with non-normal distributions.[[14]](#endnote-14) Moreover, the simplicity of the bootstrapping calculation is useful when dealing with complex distributions, distributions that have unknown properties, or problems without an established statistical calculation.[[15]](#endnote-15) These advantages lead bootstrapping to be more accurate than other prominent methodologies, including the Jackknife, in many circumstances.[[16]](#endnote-16)

However, bootstrapping is also faced with multiple drawbacks. The methodology is generally not effective in estimating population minimums or maximums, determining the sample mean when the population variance is infinite, and identifying the sample median if there is population density discontinuity at the population median.[[17]](#endnote-17) Similarly, bootstrapping does not work well with sample eigenvalues in cases where population eigenvalues have multiplicity. Another disadvantage of the technique is the necessity for a large quantity of sampling simulations that require high levels of computation, especially compared to similar approaches such as the Jackknife.[[18]](#endnote-18)

Although computational intensity hindered the spread of bootstrapping at its inception, subsequent technological and computational developments have created new opportunities to execute such demanding calculations.[[19]](#endnote-19) An especially impactful development is parallel computing, which allows computational functions to be carried out simultaneously to process high amounts of information in shorter periods of time.[[20]](#endnote-20) The technique often involves breaking large calculations into smaller ones and utilizing multiple processors to execute different functions and reduce computation time.[[21]](#endnote-21) Given the benefits of this approach, this paper examines how parallel computing can be applied to perform bootstrapping computations in an efficient and effective manner.

1. Boos, D. (2003). Introduction to the Bootstrap World. Statistical Science, 18(2), 174. Retrieved January 13, 2020, from www.jstor.org/stable/3182846 [↑](#endnote-ref-1)
2. Ibid, 168. [↑](#endnote-ref-2)
3. Ibid, 168. [↑](#endnote-ref-3)
4. Gunter, B. (1994). Bootstrapping. Infection Control and Hospital Epidemiology, 15(8), 544. Retrieved January 14, 2020, from www.jstor.org/stable/30148407 [↑](#endnote-ref-4)
5. Efron, B. (1994). The Jackknife, the bootstrap and other resampling plans, 27. Philadelphia, PA: Society for Industrial and applied mathematics. [↑](#endnote-ref-5)
6. Ibid, 27. [↑](#endnote-ref-6)
7. Ibid, 27. [↑](#endnote-ref-7)
8. Efron, Bradley. Second Thoughts on the Bootstrap. Statist. Sci. 18 (2003), no. 2, 140. doi:10.1214/ss/1063994968. https://projecteuclid.org/euclid.ss/1063994968 [↑](#endnote-ref-8)
9. Ibid, 27. [↑](#endnote-ref-9)
10. Hall, P. (1992). On Bootstrap Confidence Intervals in Nonparametric Regression. The Annals of Statistics, 20(2), 695–711. doi: 10.1214/aos/1176348652 [↑](#endnote-ref-10)
11. Politis, D. N., Romano, J. P., & Wolf, M. (1999). Subsampling. Springer Series in Statistics. doi: 10.1007/978-1-4612-1554-7 [↑](#endnote-ref-11)
12. Lahiri, S. N. (2003). Resampling Methods for Spatial Data. Resampling Methods for Dependent Data Springer Series in Statistics, 281–338. doi: 10.1007/978-1-4757-3803-2\_12 [↑](#endnote-ref-12)
13. Singh, K., & Xie, M. (2011, December 5). Bootstrap: A Statistical Method, 2. Retrieved January 16, 2020, from https://www.stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf [↑](#endnote-ref-13)
14. Ibid, 3. [↑](#endnote-ref-14)
15. Ibid, 3. [↑](#endnote-ref-15)
16. DiCiccio, T., & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science,* *11*(3), 201. Retrieved January 16, 2020, from www.jstor.org/stable/2246110 [↑](#endnote-ref-16)
17. Singh, K., & Xie, M. (2011, December 5). Bootstrap: A Statistical Method, 11. Retrieved January 16, 2020, from https://www.stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf [↑](#endnote-ref-17)
18. DiCiccio, T., & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science,* *11*(3), 201. Retrieved January 16, 2020, from www.jstor.org/stable/2246110 [↑](#endnote-ref-18)
19. Boos, D. (2003). Introduction to the Bootstrap World. Statistical Science, 18(2), 174. Retrieved January 13, 2020, from www.jstor.org/stable/3182846 [↑](#endnote-ref-19)
20. Kindervater, G., Lenstra, J., & A. H. G. Rinnooy Kan. (1989). Perspectives on Parallel Computing. Operations Research, 37(6), 985. Retrieved January 28, 2020, from www.jstor.org/stable/171480 [↑](#endnote-ref-20)
21. Ibid, 985. [↑](#endnote-ref-21)