# Deep Statistical Comparison of Meta-heuristic Stochastic Optimization Algorithms

Tome Eftimov
Computer Systems Department,
Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

Peter Korošec
Computer Systems Department,
Jožef Stefan Institute
Ljubljana, Slovenia
Faculty of Mathematics, Natural Sciences and
Information Technologies
Koper, Slovenia
peter.korosec@ijs.si

Barbara Koroušić Seljak
Computer Systems Department,
Jožef Stefan Institute
Ljubljana, Slovenia
barbara.korousic@ijs.si

## ABSTRACT

In this paper a recently proposed approach for making a statistical comparison of meta-heuristic stochastic optimization algorithms is presented. The main contribution of this approach is that the ranking scheme is based on the whole distribution, instead of using only one statistic to describe the distribution, such as average or median. Experimental results showed that our approach gives more robust results compared to state-of-the-art approaches in case when the results are affected by outliers or by statistical insignificant differences that could exist between data values.

## CCS CONCEPTS

• **Mathematics of computing** → **Hypothesis testing and confidence interval computation**;

## KEYWORDS

Statistical analysis, stochastic optimization algorithms

**ACM Reference Format:**
Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. 2018. Deep Statistical Comparison of Meta-heuristic Stochastic Optimization Algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference 2018 (GECCO '18 Companion),* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3205651.3208210

## 1 INTRODUCTION

To determine the strengths and weaknesses of a newly introduced evolutionary algorithm, its performance should be compared with performances of state-of-the-art algorithms. Over last years, several competitions for optimization algorithms at evolutionary computation conferences (e.g., GECCO, CEC) are being organized, in which the proposed algorithms are compared using a set of benchmark functions. The idea behind those comparisons is that by using the results obtained on different functions, the "best" algorithm (i.e. algorithm that perform best on average over all benchmark functions) can be found, or to use the benchmarking results to transfer the knowledge onto a real-world problem. Statistical analyses that are performed in such cases are crucial and need to be made with a great care because they provide the information from where the conclusions are made.

## 2 STATISTICAL ANALYSIS

The statistical analysis of the performance of a new algorithm with regard to state-of-the-art algorithms is in most cases made using statistical comparisons that follow the idea of hypothesis testing. Statistical comparisons can be conducted in two scenarios: single-problem analysis and multiple problem analysis. Single-problem analysis involves analyzing data from multiple runs of stochastic optimization algorithms on one problem (i.e. test function). This happens because a single run on a single problem instance is not enough to make conclusions, since the algorithms are stochastic in nature, meaning we do not have any guarantee that the result will be the same for every run; even the path leading to the final solution is often different. Multiple-problem analysis is a scenario when the algorithms are compared on a set of benchmark problems.

Nowadays, many researchers have problems making a statistical comparison because statistical tools are relatively complex and there are many to chose from. The problem is in selecting the right statistic to apply on a selected performance measure. For example, researchers often report either average or median without being aware that averaging is sensitive to outliers and both, the average and median, are sensitive to statistical insignificant differences in the data. Even reporting the standard deviation of the average needs to be made with care since large variances result from the presence of outliers. Additionally, applying the appropriate statistical test requires knowledge of the necessary conditions about the data that must be met in order to apply it. This step is often omitted and researchers simply apply a statistical test, in most cases borrowed from a similar published study, which can be inappropriate for their data set. This kind of misunderstanding is all too common in the research community and can be observed in many high-ranking journal papers. Even if the statistical test is the correct one, if the experimental design is flawed (e.g., comparison of results of tuned and non-tuned algorithms) their conclusions will be wrong. This is sometimes done on purpose to mislead the reader in believing that the author's results are better than they actually are.


*GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan*

ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/10.1145/3205651.3208210

### 2.1 Common approach

Working with stochastic optimization algorithms in multiple-problem analysis requires finding a unique representative value from multiple runs for each algorithm on each problem. For this reason, Garcia et al. [2] suggest using an average of multiple runs as a representative value for each algorithm on each problem. Average is an unbiased estimator of the expected value; however it can be affected by outliers (i.e. poor runs of stochastic optimization algorithms) and instead median can be used as a representative value. So by using the common approach, either average or median from the multiple runs obtained on a single problem can be used as a representative value involved in the multiple-problem scenario for specific algorithm on specific problem. Further the data obtained for multiple-problem analysis should be analyzed using an appropriate omnibus statistical test [2].

### 2.2 Deep statistical comparison approach

Using the common approach, we need to be aware that averages are known to be sensitive to outliers. In general, outliers can be disregarded using some techniques, but they need to be used with great care. For multiple-problem analysis, removing outliers is questionable because only the results for certain problems would be changed. In stochastic optimization it can happen that in a set of independent runs the average result of one problem for a given algorithm is better than another algorithm, but in the next set of independent runs the average result for the same problem and the same algorithm could be worse than the other algorithm, and this happens because in any new set of independent runs different poor runs exist. The common approach is also used with medians because they are less sensitive to outliers. However, in both cases the results can still be affected by the ranking scheme of some statistical tests. This happens when differences between the averages or medians are in some ε-neighborhood (e.g., $10^{-9}$,$10^{-10}$, etc.), so algorithms consequently obtain different rankings because there are no ties presented. It can happen that the distribution of the data is the same, the medians are in some ε- neigbourhood and the algorithms will be ranked differently, but they need to obtain the same ranking; even more the distribution can be different, the medians can be the same and the algorithms will be ranked as the same, but they need to obtain different rankings. All this leads to a need for new robust analyses that can be used to compose a sample for each algorithm over multiple problems, which can be used for further analysis using a standard omnibus statistical test.

For these reasons, we proposed Deep Statistical Comparison (*DSC*) for comparing meta-heuristic stochastic optimization algorithms over multiple single-objective problems [1]. Its main contribution is its ranking scheme, which is based on the whole distribution, instead of using only one statistic to describe the distribution, such as average or median. The approach consists of two steps. The first step uses a newly proposed ranking scheme to obtain data in order to make a statistical comparison. The ranking scheme is based on comparing distributions using a statistical test, such as, the two-sample *Kolmogorov-Smirnov test* or the two-sample *Anderson-Darling test*. All pairwise comparisons between the compared algorithms must be made, and the obtained p-values are organized in a matrix. Further, because multiple pairwise comparisons are made, these p-values are corrected using the *Bonferroni correction* [2] in order to control the family-wise error, FWER. The FWER is the probability of making one or more false discoveries, or type I errors, among all hypotheses when performing multiple hypotheses tests. The matrix is then checked for transitivity, and on this basis the algorithms obtain their rankings. The second step is a standard omnibus statistical test, which uses data obtained by the *DSC* ranking scheme as input.

For making a statistical comparison easier without having to worry about making incorrect conclusions, the DSC approach can be used via two HTTP REST API web services and a web-based interface developed as *Shiny* application in the *R programming language* (http://ws.ijs.si/dsc/).

## 3 DISCUSSION

Experimental results that involved comparisons of algorithms presented at the Black-Box Benchmarking 2015 competition, which was part of the GECCO 2015, showed that there are combinations of algorithms for which the common approach and the DSC approach provide different results. This happens because using the common approach with averages, averages are affected by poor runs of stochastic optimization algorithms. This could be solved by using medians. However, in both cases, either averages or medians, the problem is that they can be in some ε-neighbourhood (i.e. insignificant statistical difference) and will be ranked as different, but should be ranked as the same. All these problems can be omitted using the DSC approach, which takes into account the whole distribution of the multiple runs for an algorithm obtained on a given problem, and it is based on comparing distributions.

## 4 CONCLUSION

Working with stochastic optimization algorithms, statistical comparison plays an important role for objectively comparing new algorithms in order to demonstrate their strengths and weaknesses, where they can be improved. For this reason, in this paper we presented a recently proposed approach for making a statistical comparison of meta-heuristic stochastic optimizaiton algorithms, which provides more robust statistical results than state-of-the-art approaches when results are affected by outliers or statistical insignificant difference that could exist between data values [1].

## ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 692286 and the financial support from the Slovenian Research Agency (research core funding No. P2-0098).

## REFERENCES

Tome Eftimov, Peter Korošec, and Barbara Koroušić Seljak. 2017. A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics. *Information Sciences* 417 (2017), 186–215.

Salvador García, Daniel Molina, Manuel Lozano, and Francisco Herrera. 2009. A study on the use of non-parametric tests for analyzing the evolutionary algorithms? behaviour: a case study on the CEC2005 special session on real parameter optimization. *Journal of Heuristics* 15, 6 (2009), 617–644.