DISADVANTAGES OF STATISTICAL COMPARISON OF STOCHASTIC OPTIMIZATION ALGORITHMS

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

Peter Korošec
Computer Systems Department, Jožef Stefan Institute, Ljubljana, Slovenia
peter.korosec@ijs.si

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

Abstract In this paper a short overview and a case study in a statistical comparison of stochastic optimization algorithms are presented. The algorithms are part of the Black-Box Optimization Benchmarking 2015 competition that was held at the 5th GECCO Workshop for Real-Parameter Optimization. The question about the difference between parametric and non-parametric tests for single-problem analysis and for multiple-problem analysis is addressed in this paper. The main contributions are the disadvantages that can appear by using multiple-problem analysis, in the case when the data of some algorithms includes outliers.

Keywords: Comparative study, Non-parametric tests, Parametric tests, Statistical methods, Stochastic optimization algorithms.

1. Introduction

Over the last years, many machine learning and stochastic optimization algorithms have been developed. For each new algorithm, according

to its performance, we need to decide whether it is better than the compared algorithms used on the same problem.

One of the most common ways to compare algorithms used on the same problem is to use statistical tests as comparison techniques of their performance [4, 5, 6, 9, 10]. The common thing of the comparative studies, independently of the research area (machine learning, stochastic optimization or some other research areas), is that they are based on the idea of hypothesis testing [18].

The hypothesis testing, also called significance testing, is a method of statistical inference that could be used for testing a hypothesis about parameter in a population, using data measured in a data sample, or about the relationship between two or more populations, using data measured in data samples. The method starts by defining two hypotheses, the *null hypothesis* H_0 and the *alternative hypothesis* H_A. The null hypothesis is a statement that there is no difference or no effect and the alternative hypothesis is a statement that directly contradicts the null hypothesis by indicating the presence of a difference or an effect. This step in the hypothesis testing is very important, because mis-stating the hypotheses will disrupt the rest of the process. The second step is to select an appropriate *test statistic* T, which is a mathematical formula that allows researchers to determine the likelihood of obtaining the outcomes if the null hypothesis is true. Then, the level of significance α, also called *significance level*, which is the probability threshold below which the null hypothesis will be rejected, needs to be selected. The last step of the hypothesis testing is to make a decision either to reject the null hypothesis in favor of the alternative or not to reject it. The last step can be done with two different approaches. In the standard approach, the possible values of the test statistic for which the null hypothesis is rejected, also called the *critical region*, are calculated using the distribution of the test statistic and the probability of the critical region that is the level of significance α. Then the observed value of the test statistic T_obs is calculated according to the observations from the data sample. If the observed value of the test statistic is in the critical region, the null hypothesis is rejected, and if not, it fails to reject the null hypothesis. In the alternative approach, instead of defining the critical region, a *p-value* that is the probability of obtaining the sample outcome, given the null hypothesis is true, is calculated. The null hypothesis is rejected, if the p-value is less than the selected significance level (the most common values for it are 0.05 and 0.01), and if not, it fails to reject the null hypothesis.

In this paper we follow the recommendations given in some papers [4, 5, 6, 9, 10] in order to perform correct statistical comparison of

the behavior of some of the stochastic optimization algorithms over optimization problems presented of the Black-Box Benchmarking 2015 (BBOB 2015) competition helded at the 5th GECCO Workshop on Real-Parameter Optimization organized at the Genetic and Evolutionary Computation Conference (GECCO 2015) [1].

This paper can be seen as a tutorial and a case study on the use of statistical tests for comparison of stochastic optimization algorithms. In Section 2 we give a review and important comments with regard to the standard statistical tests, the parametric tests, and the non-parametric tests. Section 3 presents the empirical study carried out on the results from the workshop in different scenarios, using pairwise comparison for single-problem analysis, pairwise comparison for multiple-problem analysis, multiple comparisons for single-problem analysis, and multiple comparisons for multiple-problem analysis. In Section 4 we conclude the paper by discussing the disadvantages of the standard statistical tests that are used for statistical comparisons of the behavior of stochastic optimization algorithms.

2. Parametric Versus Non-parametric Statistical Tests

In order to distinguish what to use for your data, between the parametric and the non-parametric test, the first step is to check the assumptions of the parametric tests, also called required conditions for the safe use of parametric tests. So the first step is to use the methods for checking the validity of these required conditions. If the data does not satisfy the required conditions for the safe use of parametric tests, then the tests could lead to incorrect conclusions, and it is better to use the analogous non-parametric test. In general, a non-parametric test is less restrictive than a parametric one, but it is less powerful than a parametric one, when the required conditions for the safe use of the parametric test are true [10].

2.1 Required Conditions for the Safe Use of Parametric Tests

The assumptions or the required conditions for the safe use of parametric tests are the independence, the normality, and the homoscedasticity of the variances of the data.

Two events, A and B are independent, if the fact that A occurs does not affect the probability of B occurring. When we compare the behavior of the stochastic optimization algorithms, they are usually independent.

The assumption of normality is just a hypothesis that a random variable of interest, or in our case the data from the data sample, is distributed according to the normal or Gaussian distribution with mean μ and standard deviation σ. In order to check the validity of this condition, the recommended statistical tests are *Kolmogorov-Smirnov* [19], *Shapiro-Wilk* [23], and *D'Agostino-Pearson* [3]. The validity of this condition can be also checked by using graphical representation of the data using histograms and qunatile-quantile plots (Q-Q plots) [7].

The homoscedasticity indicates the hypothesis of equality of variances, and the *Levene's test* is used to check the validity of this condition [13]. Using this test we can see whether or not a given number of samples have equal variances or not.

2.2 An Overview of Some Standard Statistical Tests

In Table 1 we give an overview of the most commonly used statistical tests that can be used for statistical comparison between two or multiple algorithms. We do not go into details for each of them, because they are standard statistical tests [18]. Which of them is chosen depends on the type of analysis we want to perform, either single-problem or multiple-problem analysis.

Table 1: An overview of parametric and non-parametric tests

	Two Algorithms	*Multiple Algorithms*
Parametric tests	Paired T-Test	Repeated-Measures ANOVA
Non-parametric tests	Wilcoxon Signed-Rank Test, The Sign Test	Friedman Test, Iman-Davenport Test

The single-problem analysis is the scenario when the data comes from multiple runs of the stochastic optimization algorithms on one problem, one function. This scenario is common in stochastic optimization algorithms, since they are of stochastic nature, meaning we do not have any guaranty that the result will be optimal for every run. Moreover, typically even the path leading to the final solution is often different. So to test the quality of the algorithm, it is not enough to performed just one run, but many of them from which we can draw some conclusions.

The second scenario or the multiple-problem analysis is the scenario when several stochastic optimization algorithms are compared on multiple problems, multiple functions. In this case, in most papers the authors use the averaged results for each function to compose a sample of results for each algorithm.

3. Case Study: Black-Box Optimization Benchmarking 2015

In order to go through the recommendations of how to perform statistical comparisons of stochastic optimization algorithms and to see the possible problems that appear, the results from the Black-Box Benchmarking 2015 [1] are used. The Black-Box Benchmarking 2015 is a competition that provides single-objective functions for benchmarking. In addition, it enables analyses of the performance of the competing algorithms, and makes it understandable what are the advantages and disadvantages for each algorithm.

From the competition the algorithms *BSif*, *BSifeg*, *BSrr*, and *Srr* are used for statistical comparisons. The capital letters, S or BS, denote STEP or Brent-STEP method, respectively. The lowercase letters denote the dimension selection strategy: "rr" for round-robin, "if" for the EWMA estimate of the improvement frequency, and "ifeg" for "if" combined with ϵ-greedy strategy [21]. For each of them the results for 24 different noiseless test functions in 5 dimensionality (2, 3, 5, 10, and 20) are selected. At the end, the statistical comparison is performed by comparing the algorithms on 22 different noiseless functions because some of them do not provide data for two functions of the benchmark when the dimension is 20.

The test functions are from 5 groups: separable functions, functions with low or moderate conditioning, function with hight conditioning and unimodal, multi-modal functions with adequate global structure, and multi-modal functions with week global structure. More details about them can be found in [15].

We have done the statistical comparisons in *"R programming language"*, by using the *"lawstat" package* [11] for the *Levene's Test*, the *"stats" package* [22] for the *Kolmogorov-Smirnov Test*, the *Paired-T Test* [16], the *Shapiro-Wilk Test*, and the *Wilcoxon Signed-Rank Test* [17], and the *"scmap" package* [2] for the *Iman-Davenport Test* [6], the *Friedman Test* [6], and the *Friedman Algined-Rank Test* [6].

3.1 Pairwise and Multiple Comparisons for Single-Problem Analysis

In this section pairwise comparison for single-problem analysis is presented together with comments for multiple comparisons for single-problem analysis. The *BSif* and *BSifeg* algorithms are the two algorithms used for pairwise comparison. The pairwise comparisons between these two algorithms for single-problem analysis are performed on 22 benchmark functions when the dimension is 10.

At the beginning of each statistical comparison, the required conditions for the safe use of the parametric tests are checked.

In case of the single-problem analysis the multiple runs of the algorithm on the same function are independent.

To check for normality, the *Shapiro-Wilk Test* and graphical representations by representing the data using histograms and quantile-quantile plots are used. The p-values from the *Shapiro-Wilk Test* are presented in the Table 2, and when the p-value is smaller than the significance level (we used 0.05), then the null hypothesis is rejected, and we assume the data is not normally distributed.

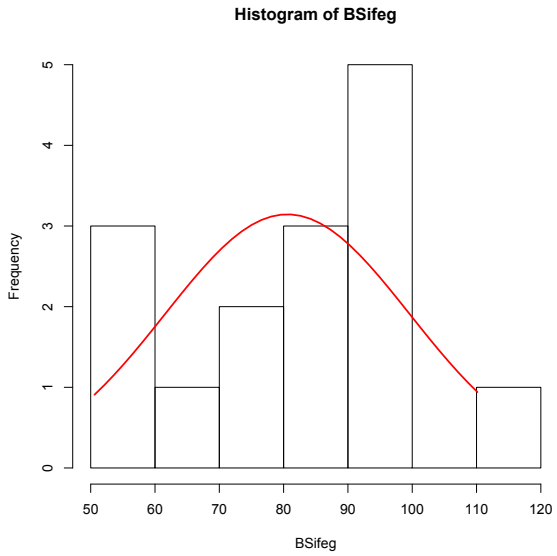
Table 2: Test of normality using *Shapiro-Wilk Test*

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
p-value BSif	-	(.61)	*(.00)	(.70)	*(.01)	*(.02)	*(.01)	(.05)
p-value BSifeg	-	*(.03)	*(.00)	(.47)	*(.01)	*(.00)	(.28)	*(.00)
	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
p-value BSif	*(.00)	*(.00)	(.49)	(.23)	*(.01)	*(.00)	*(.02)	(.05)
p-value BSifeg	*(.00)	*(.01)	(.28)	*(.00)	*(.00)	*(.00)	*(.02)	(.24)
	f_{17}	f_{18}	f_{19}	f_{20}	f_{21}	f_{22}		
p-value BSif	(.21)	*(.04)	(.16)	*(.01)	*(.00)	*(.00)		
p-value BSifeg	(.24)	(.13)	(.07)	(.10)	*(.00)	*(.00)		

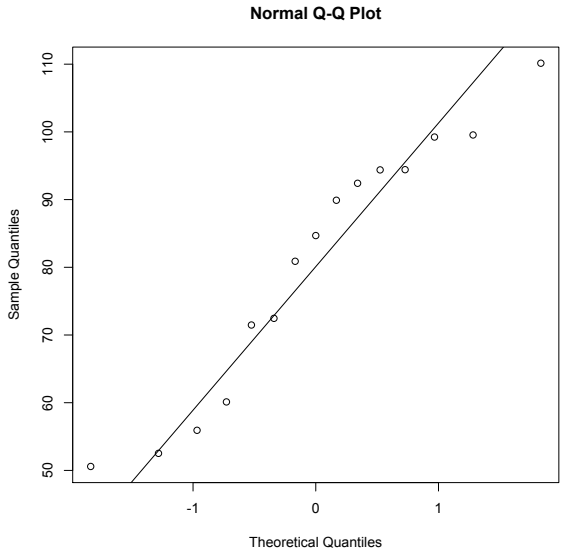
* indicates that the normality condition is not satisfied.

From the same table we can see that there are 6 cases in which the data from both algorithms comes from normal distribution (f_1, f_4, f_{11}, f_{16}, f_{17}, f_{19}), 6 cases in which the data only from one of the algorithms comes from normal distribution (f_2, f_7, f_8, f_{12}, f_{18}, f_{20}), and 10 cases in which the data from both algorithms is not normally distributed (f_3, f_5, f_6, f_9, f_{10}, f_{13},f_{14}, f_{15}, f_{21}, f_{22}).

In Fig. 1 and Fig. 2, the graphical representation of the data with normal and non-normal distribution is presented, respectively. Using the histograms, we can see if the distribution of the data is close to the shape of the distribution we are interested. The red line corresponds to the normal curve with the mean value and the standard deviation obtained from the data. The Q-Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. In our case the distribution of the data is compared with the normal distribution. If the data is normally

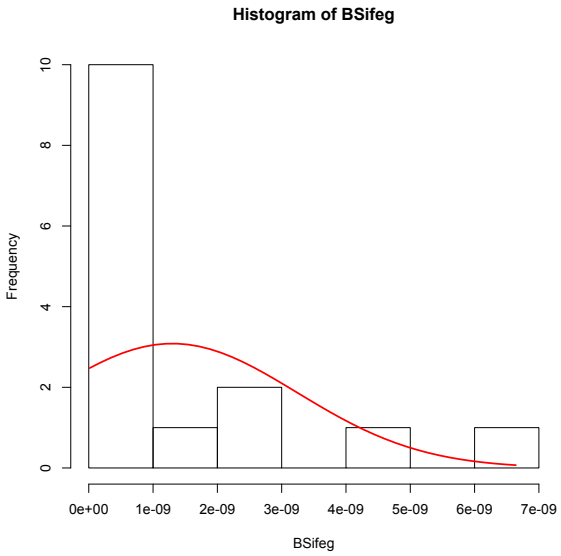


a) Histogram of *BSifeg*.

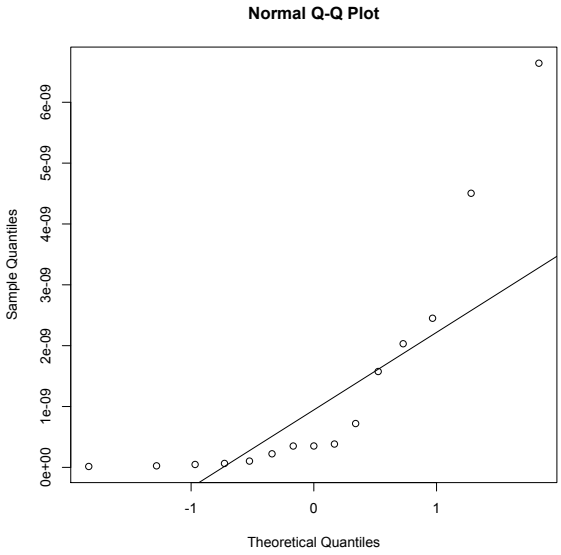


b) QQ-plot of *BSifeg*.

Figure 1: Example of normal distribution for the *BSifeg* algorithm for f_{11} with dimension 10.



a) Histogram of *BSifeg*.



b) QQ-plot of *BSifeg*.

Figure 2: Example of non-normal distribution for the *BSifeg* algorithm for f_3 with dimension 10.

distributed, the data points in the Q-Q normal plot can be approximated with a straight diagonal line.

The next step of the analysis is to check the homoscedasticity. In Table 3 the p-values from the *Levene's Test* for checking homoscedasticity, based on means, are presented. When the p-value obtained by this test is smaller than the significance level (we used 0.05), then the null hypothesis is rejected, and this indicates the existence of a violation of the hypothesis of equality of variances.

Table 3: Test of homoscedasticity using the *Levene's Test*

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
p-value	-	(.08)	(.98)	(.99)	1	(.05)	(.57)	(.07)	*(.01)
	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}
p-value	(.97)	(.37)	*(.00)	*(.00)	*(.04)	(.26)	(.27)	(.85)	(.29)
	f_{19}	f_{20}	f_{21}	f_{22}					
p-value	(.77)	(.94)	(.41)	(.66)					

* indicates that the homoscedasticity condition is not satisfied.

After checking the required conditions for the safe use of the parametric tests, the pairwise comparison of these two algorithms on each function separately is performed using the *Paired-T Test* as parametric test and the *Wilcoxon Signed-Rank Test* as non-parametric test. The p-values obtained by these two tests for the pairwise comparison of the two algorithms are presented in Table 4, where the p-value smaller than the significance level of 0.05, indicates that there is a significant statistical difference between the performance of the two algorithms.

From the Table 4 we can see that for functions f_9 and f_{22} we obtained different results according to the *Paired-T Test* and the *Wilcoxon Signed-Rank Test*. In order to select the true result, first we need to check the results for the validity of the required conditions for the safe use of parametric tests. If we look at Table 2, we can see that for these two functions the normality condition is not satisfied, so we can not use the parametric tests because they could lead to incorrect conclusions, and we need to consider the result obtained by the *Wilcoxon Signed-Rank Test*. Using this test in the case of both functions the null hypothesis is rejected using a significance level of 0.05, so there is a significant statistical difference between the performance of the two algorithms, *BSif* and *BSifeg*, over functions, f_9 and f_{22}.

Table 4: Statistical comparison of *BSif* and *BSifeg* algorithms using *Paired-T Test* and *Wilcoxon Signed-Rank Test*

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
p-value $_{\text{Paired}-\text{T}}$	-	(.19)	(.32)	(.76)	-	*(.02)	(.81)	* (.02)
p-value $_{\text{Wilcoxon}}$	-	(.08)	(.72)	(.72)	-	*(.03)	(.60)	*(.00)
	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
p-value $_{\text{Paired}-\text{T}}$	(.08)	(.40)	(.18)	(.11)	*(.00)	*(.03)	(.06)	(.64)
p-value $_{\text{Wilcoxon}}$	*(.00)	(.17)	(.26)	(.42)	*(.00)	*(.00)	(.05)	(.39)
	f_{17}	f_{18}	f_{19}	f_{20}	f_{21}	f_{22}		
p-value $_{\text{Paired}-\text{T}}$	(.34)	(.95)	(.76)	(.87)	(.14)	(.14)		
p-value $_{\text{Wilcoxon}}$	(.68)	(.71)	(.98)	(.93)	(.14)	*(.01)		

* indicates that the null hypothesis is rejected, using $\alpha = 0.05$.
p-value $_{\text{Paired}-\text{T}}$, and p-value $_{\text{Wilcoxon}}$ indicate the p-values obtained by *Paired-T Test* and *Wilcoxon Signed-Rank Test*, respectively.

If we want to perform multiple comparisons for single-problem analysis, we need to go through the same steps as in the pairwise comparison, but we need to use the *repeated-masures ANOVA* as parametric test [12], and the *Friedman Test* or *Iman-Davenport Test* as non-parametric tests. If there is significance statistical difference between the algorithms we can continue with some post-hoc procedures relevant to the test we used [6].

3.2 Pairwise and Multiple Comparisons for Multiple-Problem Analysis

In this section multiple comparisons for multiple-problem analysis are presented togehter with comments for pairwise comparison. Following the recommendations from some papers [6, 10] that addressed the same topic, the averaged results for each function with dimension 10 are used to compose a sample of results for each algorithm. The *BSifeg*, *BSrr*, and *Srr* are the algorithms used for comparison over multiple functions.

First, the conditions for the safe use of the parametric test are checked. The condition for independence is satisfied, as we explained above.

The p-values for normality condition obtained by using the *Shapiro-Wilk Test* are presented in Table 5, from where we can see that neither of the algorithms assumes that the data comes from normal distribution.

Table 5: Test of normality using *Shapiro-Wilk Test*

	BSifeg	*BSrr*	*Srr*
p-value	*(.00)	*(.00)	*(.00)

* indicates that the normality condition is not satisfied.

The homoscedasticity is checked by applying the *Levene's Test*. The p-value obtained from the *Levene's Test* is 0.63, from which it follows that the homoscedasticity is satisfied.

Because the normality condition is not satisfied, we cannot use the *repeated-measures ANOVA* as parametric test, and we can continue the analysis by using the *Friedman Test*, the *Iman-Davenport Test*, and the *Friedman Aligned-Rank Test* as non-parametric tests. The differences between these three tests and the recommendations when to use them are explained in [6], and the p-values we obtained are presented in Table 6.

Table 6: Multiple comparisons for multiple-problem analysis

	Friedman Test	*Iman-Davenport Test*	*Friedman Aligned-rank Test*
p-value	*(.03)	*(.02)	*(.04)

* indicates that the null hypothesis is rejected, using $\alpha = 0.5$.

Using the p-values reported in Table 6, according to the three tests that are used, the null hypothesis is rejected, and there is a significant statistical difference between these three algorithms.

In order to see the difference that appears between the performance of the three algorithms, the distributions of the data for each algorithm are presented in Fig. 3. In the figure we can see that there is no difference between the distributions of the data of the three algorithms that are used in multiple-problem analysis. To confirm this, we introduced the *Kolmogorov-Smirnov Test* to compare the distributions between the pairs of algorithms, and the p-values are presented in Table 7, from where we can see that the p-values obtained are greater than 0.05, so we can not reject the null hypothesis, therefor the distributions of the data between the pairs of the algorithms are the same.

The question that arises here is, if this difference between the algorithms obtained by the use of the non-parametric tests is inforced by averaging the results from multiple runs for each function to compose a sample of results for each algorithm. Averages are known to be sensitive to outliers. For example, in machine learning, different techniques that can be used to remove outliers are presented [20, 14]. It happened for

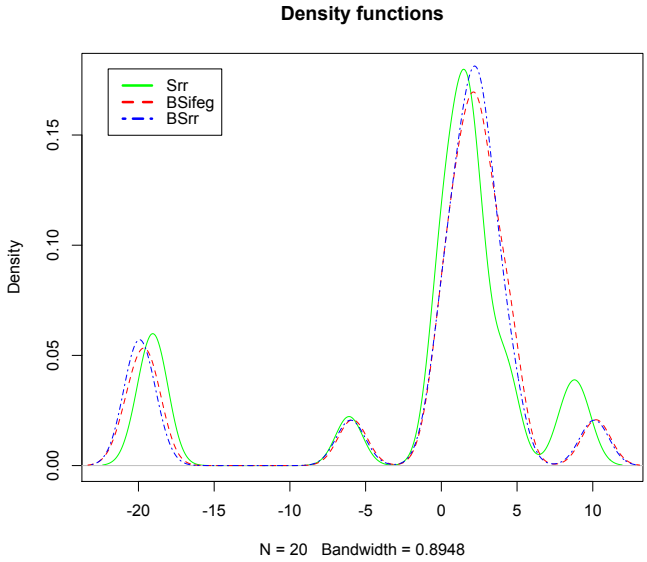


Figure 3: Probability distributions of the data of *BSifeg*, *BSrr*, and *Srr* used in multiple-problem analysis.

Table 7: Two-sample *Kolmogorov-Smirnov Test*

	(BSifeg,BSrr)	*(BSifeg,Srr)*	*(BSrr,Srr)*
p-value	1	(.86)	(.99)

example that in 15 runs the average result of one function for a given algorithm was better than another algorithm, but in new 15 runs the average result of the same function and the same algorithm could be worse than the other algorithm, and this happened because in the new 15 runs we have some outliers, or some poor runs. One solution could be to perform several multiple runs of an algorithm on the same problem, and then to average the averages results obtained by the runs. But in stochastic optimization we are not interested to have so many runs, because this is time-consuming. Another solution, also our further work, is to try to find what we can use as a measure for comparison of stochastic optimization algorithms that are robust on outliers, instead of using averaging of the results.

The pairwise comparison for multiple-problem analysis could be done using the same steps, but using the *Paired-T Tets* as parametric test,

and the *Wicoxon Signed-Rank Test* or the *The Sign Test* [8] as non-parametric tests.

4. Conclusion

In this paper a tutorial and a case study of statistical comparison between the behaviour of stochastic optimization algorithms are presented.

The main conclusion of the paper are the disadvantages that can appear in the multiple-problem analysis following the recommendations of other tutorials that address this topic. These disadvantages can happen by averaging the results from multiple runs for each function to compose a sample of results for each algorithm, in the case when the data includes outliers. In general, the outliers can be skipped using some techniques, but they need to be used with great care. But for multiple-problem analysis skipping outliers is really a question because only the results for certain problems would be changed and not for other problems. All this leads to a need of some new measures that will be robust to outliers and can be used to compose a sample for each algorithm over multiple problems, and after that to continue the analysis by using some standard statistical tests.

References

[1] Black-box benchmarking 2015. `http://coco.gforge.inria.fr/doku.php?id=bbob-2015`, accessed: 2016-02-01.

[2] B. Calvo and G. Santafe. scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal*, 2015.

[3] R. B. D'agostino, A. Belanger, and R. B. D'Agostino Jr. A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4):316–321, 1990.

[4] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[5] J. Derrac, S. García, S. Hui, P. N. Suganthan, and F. Herrera. Analyzing convergence performance of evolutionary algorithms: a statistical approach. *Information Sciences*, 289:41–58, 2014.

[6] J. Derrac, S. García, D. Molina, and F. Herrera. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18, 2011.

[7] J. Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 2015.

[8] W. J. Dixon and A. M. Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946.

[9] S. García, A. Fernández, J. Luengo, and F. Herrera. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational

intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10):2044–2064, 2010.

[10] S. García, D. Molina, M. Lozano, and F. Herrera. 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):617–644, 2009.

[11] J. L. Gastwirth, Y. R. Gel, W. L. Wallace Hui, V. Lyubchich, W. Miao, and K. Noguchi. *lawstat: Tools for Biostatistics, Public Policy, and Law*, 2015, r package version 3.0. [Online]. Available: `https://CRAN.R-project.org/package=lawstat`.

[12] E. R. Girden. *ANOVA: Repeated measures*. Sage, 1992.

[13] G. V. Glass. Testing homogeneity of variances. *American Educational Research Journal*, 3(3):187–190, 1966.

[14] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques*. Elsevier, 2011.

[15] N. Hansen, A. Auger, S. Finck, and R. Ros. Real-parameter black-box optimization benchmarking 2010: Experimental setup. Research Report RR-7215, INRIA, 2010.

[16] H. Hsu and P. A. Lachenbruch. Paired t test. In *Wiley Encyclopedia of Clinical Trials*, 2008.

[17] F. Lam and M. Longnecker. A modified wilcoxon rank sum test for paired data. *Biometrika*, 70(2):510–513, 1983.

[18] E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*. Wiley, New York, 1986.

[19] H. W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

[20] M. Nikolova. A variational approach to remove outliers and impulse noise. *Journal of Mathematical Imaging and Vision*, 20(1-2):99–120, 2004.

[21] P. Pošík and P. Baudiš. Dimension selection in axis-parallel brent-step method for black-box optimization of separable continuous functions. *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, pages 1151–1158, 2015.

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: `https://www.R-project.org/`.

[23] S. S. Shapiro and R. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216, 1972.