GECCO 2021 Tutorial: Statistical Analyses for Meta-heuristic Stochastic Optimization Algorithms

Tome Eftimov and Peter Korošec

Computer Systems Department, Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov, peter.korosec@ijs.si

https://gecco-2021.sigevo.org/

GECCO '21 Companion, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07. . . $15.00.
doi:10.1145/3449726.3461438

Instructors

Tome Eftimov is a research fellow at the Jožef Stefan Institute, Ljubljana, Slovenia. He was a postdoctoral research fellow at the Department of Biomedical Data Science, and the Centre for Population Health Sciences, Stanford University, USA, and a research associate at the University of California, San Francisco, USA (UCSF). He was awarded his PhD degree from the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, in 2018. His main areas of research include statistics, natural language processing, heuristic optimization, machine learning, and representational learning. His work related to benchmarking in computational intelligence is focused on developing more robust statistical approaches that can be used for analysis of experimental data.

Peter Korošec received his Ph.D. degree from the Jožef Stefan Postgraduate School, Ljubljana, Slovenia, in 2006. Since 2002, he has been a researcher at the Computer Systems Department, Jožef Stefan Institute, Ljubljana. He participated in organization of several conferences, workshops as either program chair or organizer. He successfully applied his optimization approaches to several real-world problems in engineering. Recently, he has focused research on better understanding of optimization algorithms so they can be more efficiently selected and applied to real-world problems.

Do we need a statistical analysis?

- Why in Evolutionary Computation?
- How good is my algorithm?
- Is it better than others?
- How good is my modeling approach?

Janez Demšar, FRI, University of Ljubljana, INIT/AERFAI Summer School on Machine Learning, Benicassim, Spain, 2015)

Tome Eftimov and Peter Korošec GECCO 2021: Statistical Analyses for Optimization Algorithms

Introduction to statistical analysis

- Descriptive statistics
  - **Distribution** of single variable
  - Measures of **central tendency**: *mean*, *median*, and *mode*
  - Measures of **variability**: *standard deviation*, *variance*, *minimum*, and *maximum*
- Inferential statistics
  - Estimation of distribution parameter
  - **Hypothesis testing**

Tome Eftimov and Peter Korošec GECCO 2021: Statistical Analyses for Optimization Algorithms

Hypothesis testing

- $\mathbf{H_0}$ "*no difference*", $\mathbf{H_A}$: "*presence of a difference*"
  - Select an appropriate omnibus statistical test
  - Calculate the value of the test statistic
  - Select the significance level
  - p-value
- Parametric vs. Nonparametric Tests
  - Independence
  - Normality
    - Kolmogorov-Smirnov, Shapiro-Wilk, D'Agostino-Pearson
  - Homoscedasticity
    - Levene's test for checking the equality of the variances

Omnibus statistical tests

	Two data samples	More than two data samples
Parametric	T-test (unpaired)	One-way ANOVA (unpaired)
	Paired t-test (paired)	Repeated-measures ANOVA (paired)
Nonparametric		Kruskal-Wallis (unpaired)
	Mann-Whitney U (unpaired)	Friedman (paired)
	Wilcoxon signed-rank (paired)	Friedman-aligned (paired)
		Iman-Davenport (paired)

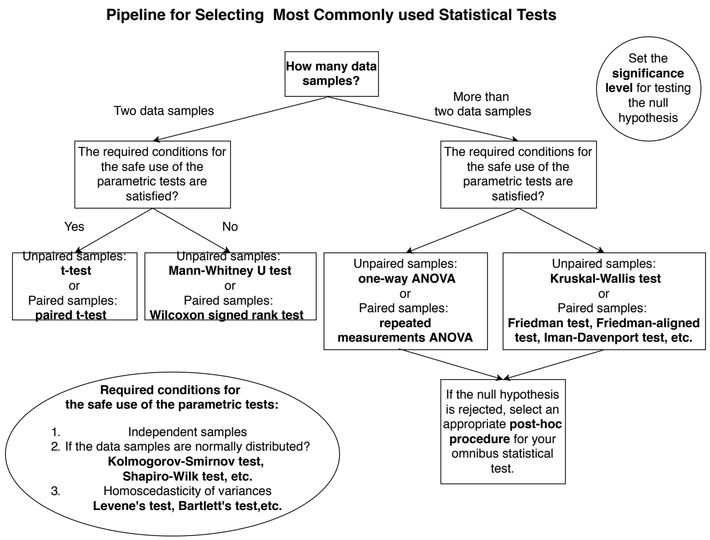
Single- vs. multiple-problem analysis

- **Single problem** (e.g., dataset, function)
  - *How good is my model?*
- **Multiple problems** (e.g., datasets, functions)
  - *How good is my modelling approach?*

Comparison scenarios

- **Pairwise Comparison**
- **Multiple Comparisons among all methods**
  - *Omnibus statistical test for more than two algorithms (e.g., Friedman test)*
    - A set of post-hoc procedures (e.g., Nemenyi, Holm, Shaffer)
- **Multiple Comparisons with a control method**
  - *Omnibus statistical test for more than two algorithms (e.g., Friedman test)*
    - A set of post-hoc procedures to compare a control method with other methods (e.g., Bonferroni-Dunn, Holland, Hochberg, Holm)
  - *Multiple pairwise tests (e.g., Wilcoxon test))*
    - The true statistical significance for combining pairwise comparisons must be calculated

A taxonomy for statistical test selection



Typical mistakes (1/2)

- **Do not borrow the statistical analysis from a similar study**
  - Check the required conditions for the safe use of the parametric tests, choose your comparison scenario and then select the appropriate statistical test for your analysis
- **Be aware how you calculate confidence intervals**
  - If your data is not normally distributed, you must calculate bootstrapping confidence intervals

Typical mistakes (2/2)

- **If you have large sample, it does not mean that it is normally distributed**
  - The central limit theorem (CLT) states that the distribution of the **sum** of a sufficiently large number of identically distributed independent random variables is approximately normal
- **Be aware that outliers exist**

Standard approaches

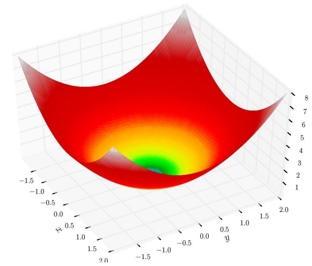
- **Demšar (Machine Learning)**
  - Demšar, J. (2006). *Statistical comparisons of classifiers over multiple data sets*. Journal of Machine learning research, 7(Jan), 1-30.
- **Garcia et al. (Evolutionary Algorithms)**
  - Derrac, J., García, S., Molina, D., & Herrera, F. (2011). *A practical tutorial 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.
  - García, S., Molina, D., Lozano, M., & Herrera, F. (2009). *A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behavior: a case study on the CEC'2005 special session on real parameter optimization*. Journal of Heuristics, 15(6), 617.

Deep Statistical Comparison

A Case Study of Meta-heuristic Stochastic Optimization Algorithms

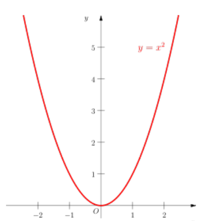
Single-objective optimization

- Single objective function



Motivation

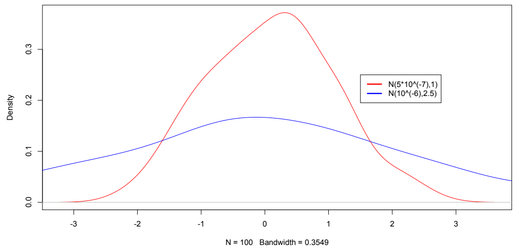
- Two stochastic optimization algorithms
  - A_1 and A_2
- 10 runs for each of them
- A_1:
  - 0;0;0;0;0;0;0;0;0;**10**
- A_2:
  - 0;1;0;0;1;1;0;1;0;1
- $mean(A_1) = 1 => ranking(A_1) = 2$
- $mean(A_2) = 0.5 => ranking(A_2) = 1$

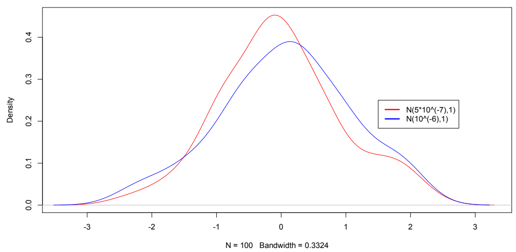


https://en.wikipedia.org/wiki/Square_(algebra)

Robust statistic

- Two stochastic optimization algorithms
  - A_1 and A_2
- 100 runs for each of them
- $A_1 : N(100, 10^{-7}, 10^{-5})$
- $A_2 : N(100, 10^{-7}, 10^{-5})$
- $mean(A_1) = 1.185 * 10^{-6} => ranking(A_1) = 2$
- $mean(A_2) = 5.558 * 10^{-7} => ranking(A_2) = 1$
- $|median(A_1) - median(A_2)| < \epsilon$





Deep Statistical Comparison (DSC)

- Deep statistics
- Two steps
  - A novel ranking scheme based on comparing distributions
  - Use an appropriate omnibus statistical test

Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2017). A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics. **Information Sciences**, 417, 186-215. Eftimov, T.,

Korošec, P., & Koroušić Seljak, B. (2018, July). Deep statistical comparison of meta-heuristic stochastic optimization algorithms. In Proceedings of the **Genetic and Evolutionary Computation Conference Companion** (pp. 15-16). ACM.

DSC ranking scheme (1/2)

- m - number of algorithms
- k - number of problems (i.e. benchmark functions)
- n - number of runs performed by each algorithm on the same problem
- Ranking is made by considering the **whole distribution**, instead of using mean or median
  - **Kruskal-Wallis** test, two-sample **Kolmogorov-Smirnov** (KS) test, **Anderson-Darling** test

DSC ranking scheme (2/2)

- $m(m-1)/2$ pairwise statistical comparisons
- M_i, a $m \times m$ matrix

$$M_i[p,q] = \begin{cases} p_{\text{value}}, & p \neq q \\ 1, & p = q \end{cases}$$

- FWER – *family-wise error rate*
  - *Bonferroni correction* test

$$M_i'[p,q] = \begin{cases} 1, & M_i[p,q] \geq \alpha_{\text{KS}}/C_m^2 \\ 0, & M_i[p,q] < \alpha_{\text{KS}}/C_m^2 \end{cases}$$

- Check the transitivity of M_i'
- Rank the algorithms

Select an appropriate omnibus statistical test

Statistical Tests	Two Algorithms	Multiple Algorithms
Parametric	paired t-test	repeated-measures ANOVA
Nonparametric	Wilcoxon signed-rank test	Friedman test, Friedman aligned-ranks test, Iman-Davenport test

Examples in single-objective optimization

- Comparison with the standard approach
- BBOB 2015 – single-objective functions for benchmarking
- 15 algorithms, 5 dimensionality (2, 3, 5, 10, and 20), 22 different noiseless functions, 15 runs for each algorithm on each function
- For the experiments we fixed the dimension on 10, D=10

Experiments

	Algorithms	Common approach		DSC approach	
		p_{valueF}	$p_{valueID}$	p_{valueF}	$p_{valueID}$
1	*RF1-CMAES, Sifeg, BSif*	*(.00)	*(.00)	(.28)	(.28)
2	*BSifeg, Sif, BSif*	*(.03)	*(.02)	(.79)	(.79)
3	*Sifeg, GP5-CMAES, BSif*	*(.02)	*(.02)	(.55)	(.56)
4	*BSif, RF1-CMAES, Sif*	*(.02)	*(.02)	(.31)	(.32)
5	*BSrr, Sif, Srr*	*(.04)	*(.03)	(.80)	(.81)
6	*RF1-CMAES, Sifeg, BSifeg*	*(.00)	*(.00)	(.07)	(.07)
7	*BSif, BSqi, BSifeg*	*(.03)	*(.03)	(.64)	(.64)
8	*BSifeg, RF1-CMAES, Srr*	*(.01)	*(.01)	(.18)	(.18)
9	*GP5-CMAES, BSif, Srr*	*(.04)	*(.04)	(.42)	(.43)
10	*BSifeg, BSrr, Srr*	*(.03)	*(.02)	(.98)	(.98)
11	*RF1-CMAES, BSifeg, Sif*	*(.02)	*(.02)	(.17)	(.17)
12	*Sifeg, GP1-CMAES, GP5-CMAES*	*(.04)	*(.04)	(.21)	(.21)
13	*Srr, BSif, BSifeg*	*(.03)	*(.03)	(.86)	(.87)
14	*Sifeg, GP5-CMAES, RF1-CMAES*	*(.03)	*(.03)	(.11)	(.11)
15	*BSrr, Sif, Sifeg*	*(.02)	*(.02)	(.77)	(.74)
16	*BSifeg, GP1-CMAES, BSqi*	(.19)	(.19)	(.70)	(.71)
17	*GP1-CMAES, BSifeg, Sif*	(.24)	(.24)	(.86)	(.87)
18	*BSqi, Srr, GP1-CMAES*	(.25)	(.26)	(.97)	(.97)
19	*BSqi, Sifeg, Sif*	(.06)	(.06)	(.86)	(.87)
20	*RF1-CMAES, BSifeg, Sif*	(.42)	(.43)	(.17)	(.17)
21	*RF1-CMAES, GP1-CMAES, Srr*	*(.00)	*(.00)	*(.02)	*(.02)
22	*BSif, Srr, CMA-CSA*	*(.00)	*(.00)	*(.00)	*(.00)
23	*BSrr, BSif, CMA-TPA*	*(.00)	*(.00)	*(.01)	*(.01)
24	*Sifeg, CMA-TPA, BSqi*	*(.00)	*(.00)	*(.00)	*(.00)
25	*BSif, BSqi, CMA-MSR*	*(.00)	*(.00)	*(.00)	*(.00)

* indicates that the null hypothesis is rejected, using $\alpha = 0.05$
p_{valueF} corresponds to the p-value obtained by the *Friedman test*
$p_{valueID}$ corresponds to the p-value obtained by the *Iman-Davenport test*

BSifeg, BSrr, and Srr (1/3)

Friedman ranking scheme (medians)

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	2.00	1.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	2.00	1.00	3.00
f_9	3.00	2.00	1.00
f_{10}	2.00	3.00	1.00
f_{11}	2.00	3.00	1.00
f_{12}	3.00	1.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	2.00	3.00	1.00
f_{20}	3.00	1.00	2.00
f_{21}	3.00	2.00	1.00
f_{22}	3.00	2.00	1.00

Friedman ranking scheme (averages)

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	1.00	2.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	1.00	2.00	3.00
f_9	3.00	2.00	1.00
f_{10}	3.00	2.00	1.00
f_{11}	1.00	3.00	2.00
f_{12}	2.00	1.00	3.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	3.00	2.00	1.00
f_{20}	3.00	2.00	1.00
f_{21}	3.00	2.00	1.00
f_{22}	2.00	3.00	1.00

DSC ranking scheme

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	1.50	1.50	3.00
f_3	1.50	1.50	3.00
f_4	2.00	2.00	2.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	2.00	2.00	2.00
f_8	2.00	2.00	2.00
f_9	2.00	2.00	2.00
f_{10}	2.00	2.00	2.00
f_{11}	2.00	2.00	2.00
f_{12}	2.00	2.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00
f_{16}	2.00	2.00	2.00
f_{17}	2.00	2.00	2.00
f_{18}	2.00	2.00	2.00
f_{19}	3.00	2.00	1.00
f_{20}	2.00	2.00	2.00
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

BSifeg, BSrr, and Srr (2/3)

Friedman ranking scheme (medians)

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	2.00	1.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	2.00	1.00	3.00
f_9	3.00	2.00	1.00
f_{10}	2.00	3.00	1.00
f_{11}	2.00	3.00	1.00
f_{12}	3.00	1.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	2.00	3.00	1.00
f_{20}	3.00	1.00	2.00
f_{21}	3.00	2.00	1.00
f_{22}	3.00	2.00	1.00

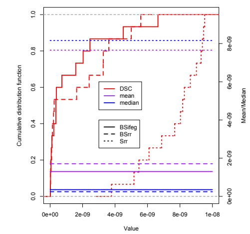
Friedman ranking scheme (averages)

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	2.00	1.00	3.00
f_3	1.00	2.00	3.00
f_4	3.00	2.00	1.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	3.00	2.00	1.00
f_8	1.00	2.00	3.00
f_9	3.00	2.00	1.00
f_{10}	3.00	2.00	1.00
f_{11}	1.00	3.00	2.00
f_{12}	2.00	1.00	3.00
f_{13}	2.00	3.00	1.00
f_{14}	3.00	2.00	1.00
f_{15}	3.00	2.00	1.00
f_{16}	2.00	3.00	1.00
f_{17}	3.00	2.00	1.00
f_{18}	2.00	3.00	1.00
f_{19}	3.00	2.00	1.00
f_{20}	3.00	2.00	1.00
f_{21}	3.00	2.00	1.00
f_{22}	2.00	3.00	1.00

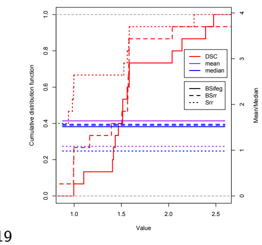
DSC ranking scheme

F	BSifeg	BSrr	Srr
f_1	1.50	1.50	3.00
f_2	1.50	1.50	3.00
f_3	1.50	1.50	3.00
f_4	2.00	2.00	2.00
f_5	2.00	2.00	2.00
f_6	2.00	3.00	1.00
f_7	2.00	2.00	2.00
f_8	2.00	2.00	2.00
f_9	2.00	2.00	2.00
f_{10}	2.00	2.00	2.00
f_{11}	2.00	2.00	2.00
f_{12}	2.00	2.00	2.00
f_{13}	2.00	3.00	1.00
f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00
f_{16}	2.00	2.00	2.00
f_{17}	2.00	2.00	2.00
f_{18}	2.00	2.00	2.00
f_{19}	3.00	2.00	1.00
f_{20}	2.00	2.00	2.00
f_{21}	2.00	2.00	2.00
f_{22}	2.00	2.00	2.00

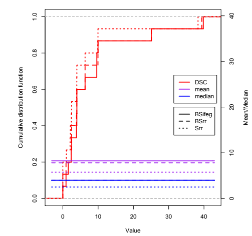
BSifeg, BSrr, and Srr (3/3)



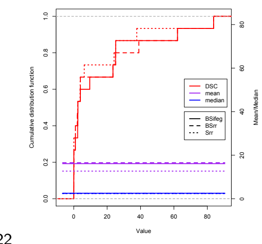
f_3



f_{19}



f_{21}



f_{22}

Using different statistical tests

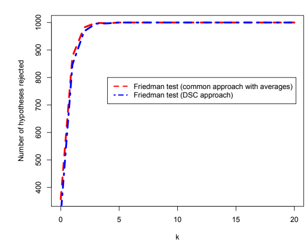
	Algorithms	Common approach	DSC approach
		p_{valueFa}	p_{valueFa}
1	*BSifeg, BSrr, Srr*	*(.00)	(.96)
2	*BSifeg, Sif, BSif*	(.37)	(.63)
3	*BSif, BSqi, CMA-MSR*	*(.00)	*(.00)

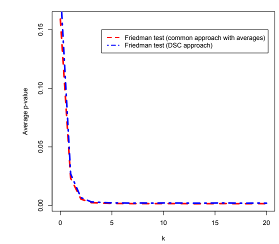
* indicates that the null hypothesis is rejected, using $\alpha = 0.05$
p_{valueFa} corresponds to the p-value obtained by the *Friedman aligned-ranks test*

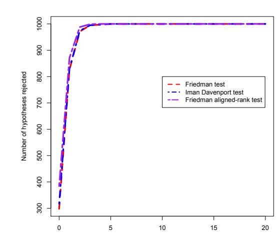
	Algorithms	Common approach		DSC approach	
		p_{value_F}	$p_{value_{ID}}$	p_{value_F}	$p_{value_{ID}}$
1	*BSqi, BSif, Srr, GP1-CMAES*	*(.04)	*(.03)	(.59)	(.60)
2	*BSqi, Sif, BSif, Srr*	*(.01)	*(.01)	(.55)	(.56)
3	*BSif, RF1-CMAES, BSifeg, GP5-CMAES*	(.60)	(.60)	(.46)	(.47)
4	*Srr, BSqi, GP1-CMAES, GP5-CMAES*	(.06)	(.06)	(.34)	(.34)
5	*BSif, CMA-MSR, BSrr, BSqi*	*(.00)	*(.00)	*(.00)	*(.00)
6	*CMA-MSR, RF1-CMAES, GP5-CMAES, BSqi*	*(.00)	*(.00)	*(.00)	*(.00)

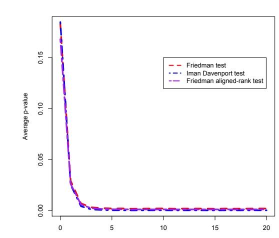
* indicates that the null hypothesis is rejected, using $\alpha = 0.05$
p_{value_F} corresponds to the p-value obtained by the *Friedman test*
$p_{value_{ID}}$ corresponds to the p-value obtained by the *Iman-Davenport test*

Power analysis



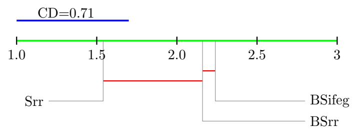




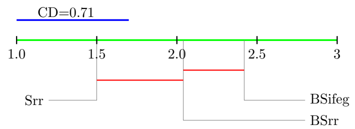


Nemenyi test

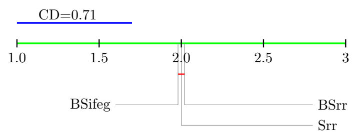
Averges



Medians



DSC



Nemenyi test

	Algorithms	Common approach	DSC approach (KS)	DSC approach (AD)
		p_{valueF}	p_{valueF}	p_{valueF}
1	*GP5-CMAES, Sifeg, BSif*	*(.02)	(.42)	(.44)
2	*BSif, RF1.CMAES, Sifeg*	*(.00)	(.28)	(.33)
3	*BSifeg, RF1-CMAES, BSrr*	(.16)	(.28)	(.48)
4	*Sif, Bsrr, GP1-CMAES*	(.35)	(.77)	(.83)
5	*BSifeg, GP1-CMAES, CMA-CSA*	*(.00)	*(.00)	*(.00)
6	*BSrr, RAND-2xDefault, Srr*	*(.00)	*(.00)	*(.00)

* indicates that the null hypothesis is rejected, using $\alpha = 0.05$
$p_{\mathrm{value\,F}}$ corresponds to the p-value obtained by the *Friedman test*

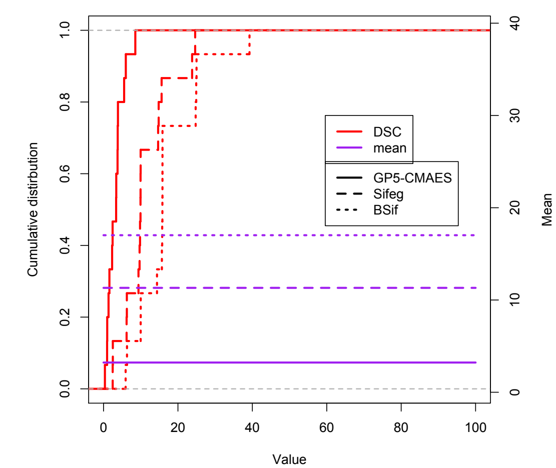
Eftimov T., Korošec P., & Koroušić Seljak, B. (2017). The Behavior of Deep Statistical Comparison Approach for Different Criteria of Comparing Distributions. In Proceedings of the **9th International Joint Conference on Computational Intelligence** - Volume 1: IJCCI, ISBN 978-989-758-274-5, pages 73-82. DOI: 10.5220/0006499900730082

Eftimov, T., & Korošec, P. (2018, July). The impact of statistics for benchmarking in evolutionary computation research. In Proceedings of the **Genetic and Evolutionary Computation Conference Companion** (pp. 1329-1336). ACM.

GP5-CMAES, Sifeg, and BSif (1/3)

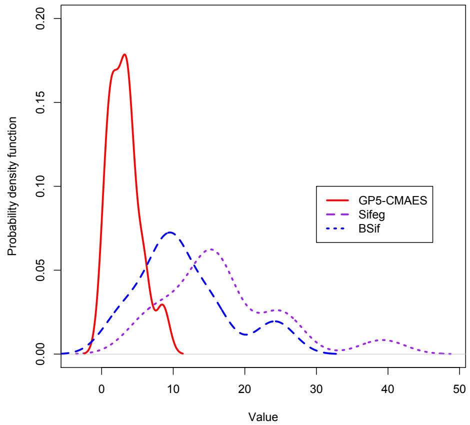
F	Common approach (*Friedman test*)			DSC ranking scheme (*KS test*)			DSC ranking scheme (*AD test*)		
	A_1	A_2	A_3	A_1	A_2	A_3	A_1	A_2	A_3
f_1	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00
f_2	3.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00
f_3	13.00	2.00	1.00	3.00	2.00	1.00	3.00	2.00	1.00
f_4	3.00	1.00	2.00	3.00	1.00	2.00	3.00	1.00	2.00
f_5	3.00	1.50	1.50	2.00	2.00	2.00	2.00	2.00	2.00
f_6	3.00	1.00	2.00	3.00	1.00	2.00	2.50	1.00	2.50
f_7	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.50	2.50
f_8	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.00	2.00
f_9	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{10}	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.50	2.50
f_{11}	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.50	2.50
f_{12}	3.00	2.00	1.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{13}	2.00	1.00	3.00	1.50	1.50	3.00	1.50	1.50	3.00
f_{14}	3.00	1.00	2.00	2.50	2.50	1.00	2.50	2.50	1.00
f_{15}	2.00	1.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00
f_{16}	2.00	1.00	3.00	2.00	2.00	2.00	2.00	1.00	3.00
f_{17}	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.50	2.50
f_{18}	1.00	2.00	3.00	1.00	2.50	2.50	1.00	2.00	3.00
f_{19}	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{20}	3.00	1.00	2.00	3.00	1.50	1.50	3.00	1.50	1.50
f_{21}	1.00	2.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00
f_{22}	1.00	2.00	3.00	2.00	2.00	2.00	2.00	2.00	2.00

GP5-CMAES, Sifeg, and BSif (2/3)



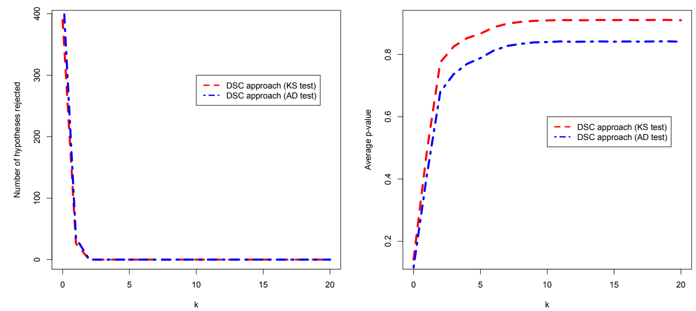
f_{18}

GP5-CMAES, Sifeg, and BSif (3/3)



f_{18}

Power analysis



Multiple comparison with a control algorithm

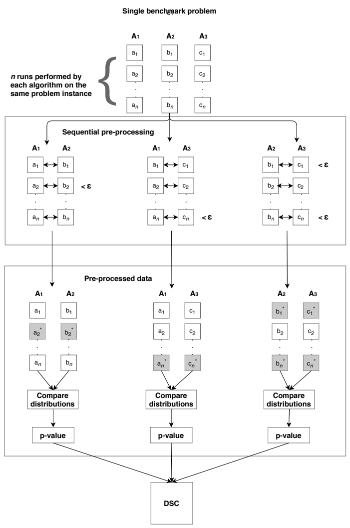
$$p_{value} = 1 - \prod_{i=1}^{k-1} [1 - p_{value_{H_i}}]$$

j	CMA-CSA vs.	$p-value_{(DSC;KS)}$	$p-value_{(DSC;AD)}$	$p-value_{(CA;average)}$	$p-value_{(CA;median)}$
1	BSif	4.847534e-03	4.847534e-03	8.476892e-04	8.476892e-04
2	BSifeg	7.768118e-03	7.768118e-03	1.086096e-03	1.758873e-03
3	BSqi	3.081757e-03	7.768118e-03	1.227287e-03	2.223195e-03
4	BSrr	7.768118e-03	7.768118e-03	1.086096e-03	1.758873e-03
5	CMA-MSR	1.000000e+00	7.655945e-01	4.757041e-02	7.628835e-02
6	CMA-TPA	1.000000e+00	3.457786e-01	4.757041e-02	4.654448e-01
7	GP1-CMAES	1.451271e-05	8.553503e-06	4.768372e-07	6.411516e-05
8	GP5-CMAES	8.553503e-06	8.553503e-06	4.768372e-07	6.411516e-05
9	RAND-2xDefault	5.049088e-06	5.049088e-06	6.411516e-05	6.411516e-05
10	RF1-CMAES	5.049088e-06	5.049088e-06	4.768372e-07	6.411516e-05
11	RF5-CMAES	5.049088e-06	5.049088e-06	4.768372e-07	6.411516e-05
12	Sif	6.301490e-04	3.759531e-04	9.600603e-04	1.385265e-03
13	Sifeg	6.301490e-04	6.301490e-04	1.385265e-03	3.504330e-03
14	Srr	1.056542e-03	6.301490e-04	1.227287e-03	2.495261e-03

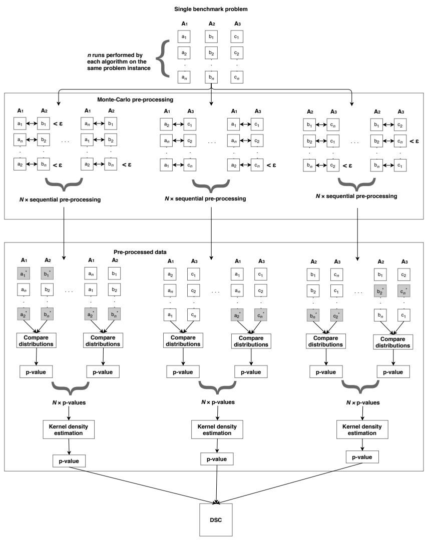
Practical Deep Statistical Comparison (pDSC)

- Practical significance instead of statistical significance
- Pre-processing for practical significance
- Sequential pDSC and Monte-Carlo ranking scheme
- Eftimov, T., & Korošec, P. (2019). Identifying practical significance through statistical comparison of meta-heuristic stochastic optimization algorithms. **Applied Soft Computing**, 105862.

Sequential pDSC



Monte-Carlo pDSC



Benchmarking for practical significance (1/3)

Statistical comparisons of 3 algorithms.

	Algorithms	pDSC approach	CRS4EAs
1	*RF1-CMAES, Sifeg, BSif*	1	0
2	*BSifeg, Sif, BSif*	1	0
3	*Sifeg, GP5-CMAES, BSif*	1	0
4	*BSif, RF1-CMAES, Sif*	1	0
5	*BSrr, Sif, Srr*	1	1
6	*RF1-CMAES, Sifeg, BSifeg*	1	0
7	*BSif, BSqi, BSifeg*	1	1
8	*BSifeg, RF1-CMAES, Srr*	1	0
9	*GP5-CMAES, BSif, Srr*	1	1
10	*BSifeg, BSrr, Srr*	1	1
11	*RF1-CMAES, BSifeg, Sif*	1	1
12	*Sifeg, GP1-CMAES, GP5-CMAES*	1	0
13	*Srr, BSif, BSifeg*	1	0
14	*Sifeg, GP5-CMAES, RF1-CMAES*	1	0
15	*BSrr, Sif, Sifeg*	1	1
16	*BSifeg, GP1-CMAES, BSqi*	1	1
17	*GP1-CMAES, BSifeg, Sif*	1	1
18	*BSqi, Srr, GP1-CMAES*	1	1
19	*BSqi, Sifeg, Sif*	1	0
20	*RF1-CMAES, BSifeg, Sif*	1	1
21	*RF1-CMAES, GP1-CMAES, Srr*	0	0
22	*BSif, Srr, CMA-CSA*	0	0
23	*BSrr, BSif, CMA-TPA*	0	0
24	*Sifeg, CMA-TPA, BSqi*	0	0
25	*BSif, BSqi, CMA-MSR*	0	0

0 - means that there is a statistical significance between the performance of the algorithms.
1 - means that there is no statistical significance between the performance of the algorithms.

Benchmarking for practical significance (2/3)

Statistical comparison of three algorithms using *CRS4EAs*.

	Algorithms	ϵ_d							
		10^{-9}	10^{-6}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
1	*RF1-CMAES, Sifeg, BSif*	0	0	0	0	0	0	1	1
2	*BSifeg, Sif, BSif*	1	0	0	0	0	0	1	1
3	*Sifeg, GP5-CMAES, BSif*	0	0	0	0	0	0	1	1
4	*BSif, RF1-CMAES, Sif*	0	0	0	0	0	0	1	1
5	*BSrr, Sif, Srr*	1	0	0	0	0	1	1	1
6	*RF1-CMAES, Sifeg, BSifeg*	0	0	0	0	0	0	1	1
7	*BSif, BSqi, BSifeg*	1	1	1	1	1	1	1	1
8	*BSifeg, RF1-CMAES, Srr*	0	0	0	0	0	0	1	1
9	*GP5-CMAES, BSif, Srr*	0	0	0	0	0	0	1	1
10	*BSifeg, BSrr, Srr*	1	0	0	0	0	0	1	1
11	*RF1-CMAES, BSifeg, Sif*	0	0	0	0	0	0	1	1
12	*Sifeg, GP1-CMAES, GP5-CMAES*	0	0	0	0	1	1	1	1
13	*Srr, BSif, BSifeg*	0	0	0	0	0	0	1	1
14	*Sifeg, GP5-CMAES, RF1-CMAES*	0	0	0	0	0	1	1	1
15	*BSrr, Sif, Sifeg*	1	1	1	1	1	1	1	1
16	*BSifeg, GP1-CMAES, BSqi*	1	1	1	1	1	1	1	1
17	*GP1-CMAES, BSifeg, Sif*	1	1	1	1	1	1	1	1
18	*BSqi, Srr, GP1-CMAES*	1	1	1	1	1	1	1	1
19	*BSqi, Sifeg, Sif*	1	1	1	1	1	1	1	1
20	*RF1-CMAES, BSifeg, Sif*	0	0	0	0	0	0	1	1
21	*RF1-CMAES, GP1-CMAES, Srr*	0	0	0	0	0	0	0	1
22	*BSif, Srr, CMA-CSA*	0	0	0	0	0	0	0	1
23	*BSrr, BSif, CMA-TPA*	0	0	0	0	0	0	0	1
24	*Sifeg, CMA-TPA, BSqi*	0	0	0	0	0	0	1	1
25	*BSif, BSqi, CMA-MSR*	0	0	0	0	0	0	0	1

0 - indicates that there is a statistical significance between the performance of the algorithms using the 95% *RI*.
1 - indicates that there is no statistical significance between the performance of the algorithms using the 95% *RI*.

Benchmarking for practical significance (3/3)

Statistical comparison of three algorithms using the sequential variant of the *pDSC*.

	Algorithms	ϵ_p							
		10^{-9}	10^{-6}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}
1	*RF1-CMAES, Sifeg, BSif*	1	1	1	1	1	1	1	1
2	*BSifeg, Sif, BSif*	1	1	1	1	1	1	1	1
3	*Sifeg, GP5-CMAES, BSif*	1	1	1	1	1	1	1	1
4	*BSif, RF1-CMAES, Sif*	1	1	1	1	1	1	1	1
5	*BSrr, Sif, Srr*	1	1	1	1	1	1	1	1
6	*RF1-CMAES, Sifeg, BSifeg*	1	1	1	1	1	1	1	1
7	*BSif, BSqi, BSifeg*	1	1	1	1	1	1	1	1
8	*BSifeg, RF1-CMAES, Srr*	1	1	1	1	1	1	1	1
9	*GP5-CMAES, BSif, Srr*	1	1	1	1	1	1	1	1
10	*BSifeg, BSrr, Srr*	1	1	1	1	1	1	1	1
11	*RF1-CMAES, BSifeg, Sif*	1	1	1	1	1	1	1	1
12	*Sifeg, GP1-CMAES, GP5-CMAES*	1	1	1	1	1	1	1	1
13	*Srr, BSif, BSifeg*	1	1	1	1	1	1	1	1
14	*Sifeg, GP5-CMAES, RF1-CMAES*	1	1	1	1	1	1	1	1
15	*BSrr, Sif, Sifeg*	1	1	1	1	1	1	1	1
16	*BSifeg, GP1-CMAES, BSqi*	1	1	1	1	1	1	1	1
17	*GP1-CMAES, BSifeg, Sif*	1	1	1	1	1	1	1	1
18	*BSqi, Srr, GP1-CMAES*	1	1	1	1	1	1	1	1
19	*BSqi, Sifeg, Sif*	1	1	1	1	1	1	1	1
20	*RF1-CMAES, BSifeg, Sif*	1	1	1	1	1	1	1	1
21	*RF1-CMAES, GP1-CMAES, Srr*	0	0	0	1	1	1	1	1
22	*BSif, Srr, CMA-CSA*	0	0	0	0	0	0	1	1
23	*BSrr, BSif, CMA-TPA*	0	0	0	0	0	0	1	1
24	*Sifeg, CMA-TPA, BSqi*	0	0	0	0	0	0	1	1
25	*BSif, BSqi, CMA-MSR*	0	0	0	0	0	0	1	1

0 - indicates that the null hypothesis is rejected, $p_{value} < 0.05$.
1 - indicates that the null hypothesis fails to reject, $p_{value} \geq 0.05$.
p_{value} corresponds to the *p*-value obtained by the *Friedman test*.

Extended Deep Statistical Comparison (eDSC)

Statistical significance ?		Distribution of the solutions in the search space	
		Yes	**No**
Distribution of the obtained solutions values (DSC)	**Yes**	The poorer performing algorithm **lacks exploration power**, while its **exploitation power cannot be assessed** and therefore, first, the exploration power needs to be improved	The compared algorithms are able to find a region with good solutions (have the **same exploration power**), but one is able to find statistically better solutions than the other in the same region (have **different exploitation powers**)
	No	The algorithm with the sparser distribution of obtained solutions has **better exploration power**	The compared algorithms have the **same exploration** and **exploitation power**

eDSC Ranking scheme (1/2)

- eDSC ranking scheme
- m - number of algorithms
- k - number of problems (i.e. benchmark functions)
- n - number of independent runs performed by each algorithm on the same problem
- d - dimension of the search space (i.e. $d >= 2$)
- Comparing distributions in high-dimensional space
  - Multivariate ϵ test

Eftimov, T., & Korošec, P. (2019). A novel statistical approach for comparing meta-heuristic stochastic optimization algorithms according to the distribution of solutions in the search space. **Information Sciences**, 489, 255-273.

Eftimov, T., & Korošec, P. (2019, July). Understanding exploration and exploitation powers of meta-heuristic stochastic optimization algorithms through statistical analysis. In Proceedings of the **Genetic and Evolutionary Computation Conference Companion** (pp. 21-22). ACM.

eDSC Ranking scheme (2/2)

- All pairwise comparison should be done

$$N_i[p,q] = \begin{cases} p_{\text{value}}, & p \neq q \\ 1, & p = q \end{cases}, \qquad N'_i[p,q] = \begin{cases} 1, & N_i[p,q] \geq \alpha_X / C_m^2 \\ 0, & N_i[p,q] < \alpha_X / C_m^2 \end{cases}.$$

  i is the problem and p, $q = 1, \ldots, m$.
- Check for transitivity of N'_i
- Cluster or sparse solutions
- Measure for multivariate spread

$$V_{i,l} = \sqrt{det(\Sigma_{i,l})} = \prod_{d_i=1}^{d} \sqrt{\lambda_{d_i}},$$

Experiments and Results

- BBOB 2009 – single-objective functions for benchmarking
- 17 algorithms out of 32 algorithms
- 5 dimensionality (2, 3, 5, 10, and 20)
- 22 different noiseless functions
- 15 runs for each algorithm on each function
- For the experiments, we fixed the dimension on 2 and 10

Statistical Comparison of Three Algorithms

Statistical comparisons of 3 algorithms.

	Algorithms	p_{valueY}	p_{valueX}
1	*Cauchy-EDA, MCS, iAMALGAM*	(.44)	(.95)
2	*FULLNEWUOA, ALPS, Cauchy-EDA*	(.47)	(.95)
3	*Cauchy-EDA, AMALGAM, MCS*	(.33)	(.95)
4	*PSO_Bounds, Cauchy-EDA, POEMS*	(.98)	(.97)
5	*Cauchy-EDA, Rosenbrock, EDA-PSO*	(.57)	(.95)
6	*EDA-PSO, LSstep, Rosenbrock*	*(.00)	(.87)
7	*PSO_Bounds, LSfminbnd, MCS*	*(.00)	(.86)
8	*LSfminbnd, VNS, iAMALGAM*	*(.00)	(.97)
9	*GA, LSstep, G3PCX*	*(.00)	(.87)
10	*G3PCX, Rosenbrock, LSfminbnd*	*(.00)	(.87)

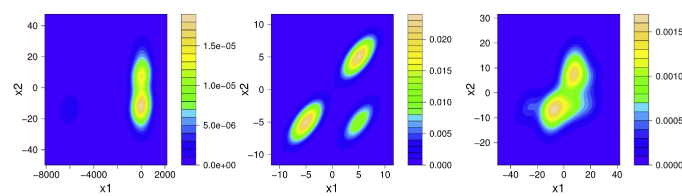
* Indicates that the null hypothesis is rejected, using $\alpha = 0.05 p_{\text{value Y}}$ corresponds to the p-value for comparing the obtained solutions values by the *Friedman test* $p_{\text{value X}}$ corresponds to the p-value for comparing distributions of the obtained solutions in search space by the *Friedman test*

Rankings on Single Problem Level

Rankings for the algorithms Cauchy-EDA, MCS, and iAMALGAM.

(a) *DSC ranking scheme*				(b) *eDSC ranking scheme*			
F	Cauchy-EDA	MCS	iAMALGAM	*F*	Cauchy-EDA	MCS	iAMALGAM
f_1	2.50	1.00	2.50	f_1	2.00	2.00	2.00
f_2	1.50	3.00	1.50	f_2	2.00	2.00	2.00
f_3	2.50	1.00	2.50	f_3	2.00	2.00	2.00
f_4	3.00	1.50	1.50	f_4	2.00	2.00	2.00
f_5	2.00	2.00	2.00	f_5	3.00	1.00	2.00
f_6	2.00	2.00	2.00	f_6	2.00	2.00	2.00
f_7	2.00	2.00	2.00	f_7	2.00	2.00	2.00
f_8	2.50	1.00	2.50	f_8	2.00	2.00	2.00
f_9	2.50	1.00	2.50	f_9	2.00	2.00	2.00
f_{10}	1.50	3.00	1.50	f_{10}	2.00	2.00	2.00
f_{11}	1.50	3.00	1.50	f_{11}	2.00	2.00	2.00
f_{12}	2.00	2.00	2.00	f_{12}	2.00	2.00	2.00
f_{13}	1.50	3.00	1.50	f_{13}	2.00	2.00	2.00
f_{14}	1.50	3.00	1.50	f_{14}	2.00	2.00	2.00
f_{15}	2.00	2.00	2.00	f_{15}	2.00	2.00	2.00
f_{16}	1.50	3.00	1.50	f_{16}	2.00	2.00	2.00
f_{17}	2.50	1.00	2.50	f_{17}	2.00	2.00	2.00
f_{18}	2.50	1.00	2.50	f_{18}	2.00	2.00	2.00
f_{19}	3.00	1.50	1.50	f_{19}	2.00	2.00	2.00
f_{20}	3.00	1.00	2.00	f_{20}	2.00	2.00	2.00
f_{21}	2.50	1.00	2.50	f_{21}	2.00	2.00	2.00
f_{22}	3.00	1.00	2.00	f_{22}	2.00	2.00	2.00

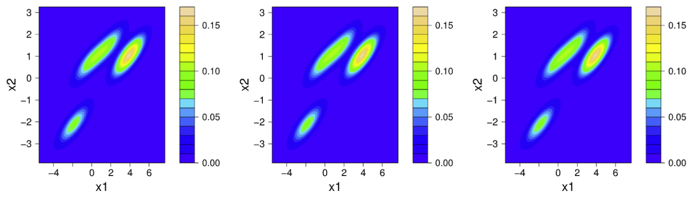
f_5 – (Cauchy–EDA, MCS, iAMALGAM)



(a) Cauchy-EDA (b) MCS (c) iAMALGAM

Fig. 3. Contour plots for probability density functions of the obtained 2-dimensional solutions for each algorithm on f_5.

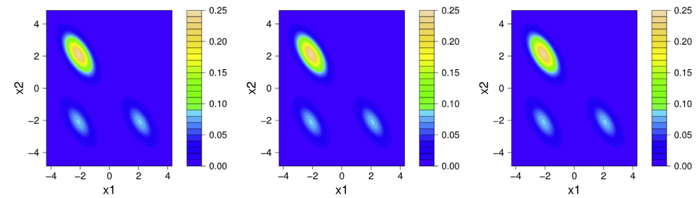
f_7 – (Cauchy–EDA, MCS, iAMALGAM)



(a) Cauchy-EDA (b) MCS (c) iAMALGAM

Fig. 6. Contour plots for probability density functions of the obtained 2-dimensional solutions for each algorithm on f_7.

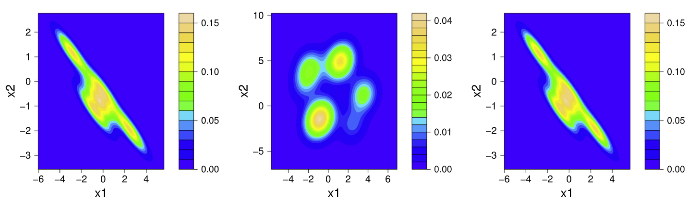
f_{20} – (Cauchy–EDA, MCS, iAMALGAM)



(a) Cauchy-EDA (b) MCS (c) iAMALGAM

Fig. 9. Contour plots for probability density functions of the obtained 2-dimensional solutions for each algorithm on f_{20}.

f_8 – (PSO-Bounds, LSfminbnd, MCS)



(a) PSO_Bounds (b) LSfminbnd (c) MCS

Fig. 12. Contour plots for probability density functions of the obtained 2-dimensional solutions for each algorithm on f_8.

Benefits of using eDSC

- Deeper understanding of exploration and exploitation powers of the compared algorithms
- Application to large-scale continuous optimization
- How the search ability of the algorithms depends on either their initial conditions or on the parameters?

How to use DSC?

- RESTful Web Services
- https://ws.ijs.si:8443/dsc-1.5/documentation.pdf
- Eftimov, T., Petelin, G., & Korošec, P. (2020). DSCTool: a web-service-based framework for statistical comparison of stochastic optimization algoritms. **Applied Soft Computing**, Volume 87, DOI: 10.1016/j.asoc.2019.105977.

DSCTool Demo - registration

- Usage: register so the rest of web services can be accessed.
- Call the service with your information inserted

```
curl -X POST
    -H 'Content-Type: application/json'
    'https://ws.ijs.si:8443/dsc-1.5/service/manage/user'
    -d '{
        "name": "YOUR NAME",
        "affiliation": "AFFILIATION INFO",
        "email": "EMAIL@ADDRESS",
        "username": "USERNAME",
        "password": "PASSWORD"
    }'
```


DSCTool DEMO - examples

- Input jsons for the examples in next slides can be accessed at http://cs.ijs.si/dl/dsctool/DSC-tutorial.zip
- Single objective (in folder "so")
  - rank_so.json
  - multivariate.json
  - omnibus_r.json
  - omnibus_m.json
  - posthoc_r.json

DSCTool Demo - rank web service

- Usage: acquire robust algorithms ranking
- For multi objective, calculate desired performance measure (e.g., quality indicator)
- Decide on statistics method (KS, AD) and significance level (alpha)
- Prepare appropriate input json and call the web service

```
curl -u 'USERNAME'
    -X POST
    -H 'Content-Type: application/json'
    'https://ws.ijs.si:8443/dsc-1.5/service/rank'
    -d@rank_so.json
```

DSCTool Demo - multivariate web service

- Usage: acquire robust algorithms ranking with respect to distribution of solutions
- Decide on desired distribution (clustered, spread) and significance level (alpha)
- Prepare appropriate input json and call the web service

```
curl -u 'USERNAME'
    -X POST
    -H 'Content-Type: application/json'
    'https://ws.ijs.si:8443/dsc-1.5/service/multivariate'
    -d@multivariate.json
```


DSCTool Demo - omnibus web service

- Usage: identify if there are any significant differences between the ranked algorithms
- Take results from any of the ranking based web services
- Select one of provided suitable omnibus tests and decide on significance level (alpha)
- Prepare appropriate input json and call the web service

```
curl -u 'USERNAME'
    -X POST
    -H 'Content-Type: application/json'
    'https://ws.ijs.si:8443/dsc-1.5/service/omnibus'
    -d@omnibus_{r|m}.json
```


DSCTool Demo - posthoc web service

- Usage: if omnibus test identifies significant differences, check from where this differences comes from
- Take results from omnibus web service
- Decide on control algorithm
- Prepare appropriate input json and call the web service

```
curl -u 'USERNAME'
    -X POST
    -H 'Content-Type: application/json'
    'https://ws.ijs.si:8443/dsc-1.5/service/posthoc'
    -d@posthoc_r.json
```


In general ...

- Novel approach for multi-target regression evaluation and multi-label classification evaluation
  - Eftimov, T., & Kocev, D. (2019). Performance Measures Fusion for Experimental Comparison of Methods for Multi-label Classification. In **AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering**
- Novel approach for benchmarking recommender systems
  - Paudel, P., Kocev, D., & Eftimov, T. (2019). Mix and Rank: A Framework for Benchmarking Recommender Systems. In Proceedings of **IEEE BigData 2019**
- Multi-level information fusion for learning a blood pressure predictive model using sensor data
  - Simjanoska, M., Kochev, S., Tanevski, J., Bogdanova, A. M., Papa, G., & Eftimov, T. (2020) In **Information Fusion**

Conclusion

In my own development (of special relativity theory) Michelson's result has not had a considerable influence. I even do not remember if I knew of it at all when I write my first paper on the subject (1905). The explanation is that I was, for general reasons, firmly convinced that there does not exist absolute motion and my problem was only how this could be reconciled with our knowledge of electro-dynamics.

Albert Einstein

References

- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan), 1-30.
- García, S., Molina, D., Lozano, M., & Herrera, F. (2009). A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization. Journal of Heuristics, 15(6), 617.
- Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2017). A novel approach to statistical comparison of meta-heuristic stochastic optimization algorithms using deep statistics. Information Sciences, 417, 186-215.
- Eftimov T., Korošec P., & Koroušić Seljak, B. (2017). The Behavior of Deep Statistical Comparison Approach for Different Criteria of Comparing Distributions. In Proceedings of the 9th International Joint Conference on Computational Intelligence - Volume 1: IJCCI, ISBN 978-989-758-274-5, pages 73-82. DOI: 10.5220/0006499900730082.
- Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2017, September). Deep Statistical Comparison Applied on Quality Indicators to Compare Multi-objective Stochastic Optimization Algorithms. In International Workshop on Machine Learning, Optimization, and Big Data (pp. 76-87). Springer, Cham.
- Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2018, December) Comparing Multi-Objective Optimization Algorithms Using an Ensemble of Quality Indicators with Deep Statistical Comparison Approach. In Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI 2017)

References

- Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2018, July). Deep statistical comparison of meta-heuristic stochastic optimization algorithms. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (pp. 15-16). ACM.
- Eftimov, T., & Korošec, P. (2018, July). The impact of statistics for benchmarking in evolutionary computation research. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (pp. 1329-1336). ACM.
- Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2018, May). Data-Driven Preference-Based Deep Statistical Ranking for Comparing Multi-objective Optimization Algorithms. In International Conference on Bioinspired Methods and Their Applications (pp. 138-150). Springer, Cham.
- Eftimov, T., & Korošec, P. (2019). Identifying practical significance through statistical comparison of meta-heuristic stochastic optimization algorithms. Applied Soft Computing, 105862.
- Eftimov, T., & Korošec, P. (2019). A novel statistical approach for comparing meta-heuristic stochastic optimization algorithms according to the distribution of solutions in the search space. Information Sciences, 489, 255-273.
- Eftimov, T., & Korošec, P. (2019, July). Understanding exploration and exploitation powers of meta-heuristic stochastic optimization algorithms through statistical analysis. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (pp. 21-22). ACM.
- Eftimov, T., Petelin, G., & Korošec, P. (2020). DSCTool: a web-service-based framework for statistical comparison of stochastic optimization algoritms. Applied Soft Computing, Volume 87, ISSN 1568-4946,

Acknowledgments

- Slovenian Research Agency - P2-0098
  - http://cs.ijs.si/
  - Follow us on Twitter: @JSI_CompSystems
- Slovenian Research Agency - Z2-1867
  - MrBec: Modern approaches for Benchmarking in Evolutionary Computation
  - http://cs.ijs.si/project/mrbec/

MrBEC