Case Study 02: Analysis of Algorithm Parameters

July, 2019

Jair Pereira Junior, 201920747

Experiment Design for Computer Science
Department of Computer Science
Graduate School of Systems and Information Engineering
University of Tsukuba

1 Introduction

Many algorithms in computer science have parameters that influence their behavior. The effect of these parameters on the algorithms can be estimated using the experiment design and analysis techniques studied in this course. In this final case study, we verify if there is significant difference in performance of four recombination operators for the algorithm Differential Evolution (implemented using the ExpDE R package). The recombination operators considered in this study are: Arithmetic, Binomial, BLX-alpha, and Eigenvector-based binomial.

The code and data is published in our **Github repository**, including the scripts to gather, visualize, and process the data. [1].

2 Pilot Study

A pilot study was conduct to determine the statistical test to be used and its required sample size. For this, a sample size $n_{pilot} = 100$ was used. Figure 1 shows the boxplot of the data gathered. As shown, there may be a difference among the performance of the recombination operators. Although there are some outliers, the data seem to have little or no skew, indicating that the data may be normal. In addiction, the operator Eigenvector-based binomial presents a huge variance, it may indicate that the data does not present homoscedasticity.

In order to use a parametric test, the data was checked for normality and for homogeneity of variance. First, to assure normality, QQ-Plot and the Shapiro test on the residuals was used. The Shapiro test indicates normality, with p-value 0.145. Figure 2-a shows the QQ-plot. As can be seen, the data seem to be normal, besides some indications that the data may be heavy-tailed. Next, to assure homoscedasticity, the Fligner-Killeen test and the plot of residuals by fitted values is shown. The Fligner-Killeen test did not reject the null-hypothesis with p-value = 2.9e-13, indicating that the data may not have homogeneity of variances. Figure 2-b shows the plot of residuals by fitted values. As shown, there is a operator with huge variance compared to the others. This operator is the Eigenvector-based binomial, and it can be identified by the boxplot on figure 1. Despite this operator, the others three seem to have similar variance.

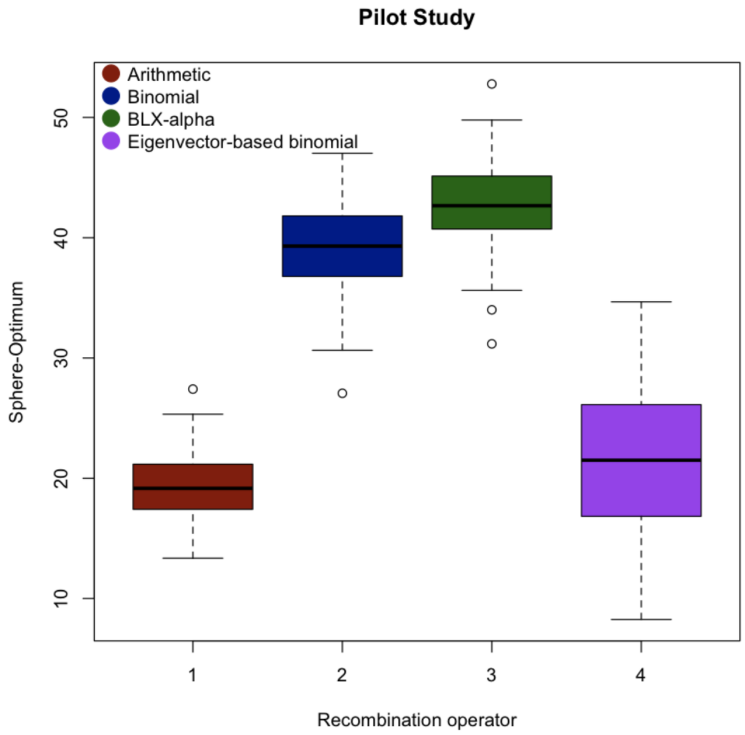
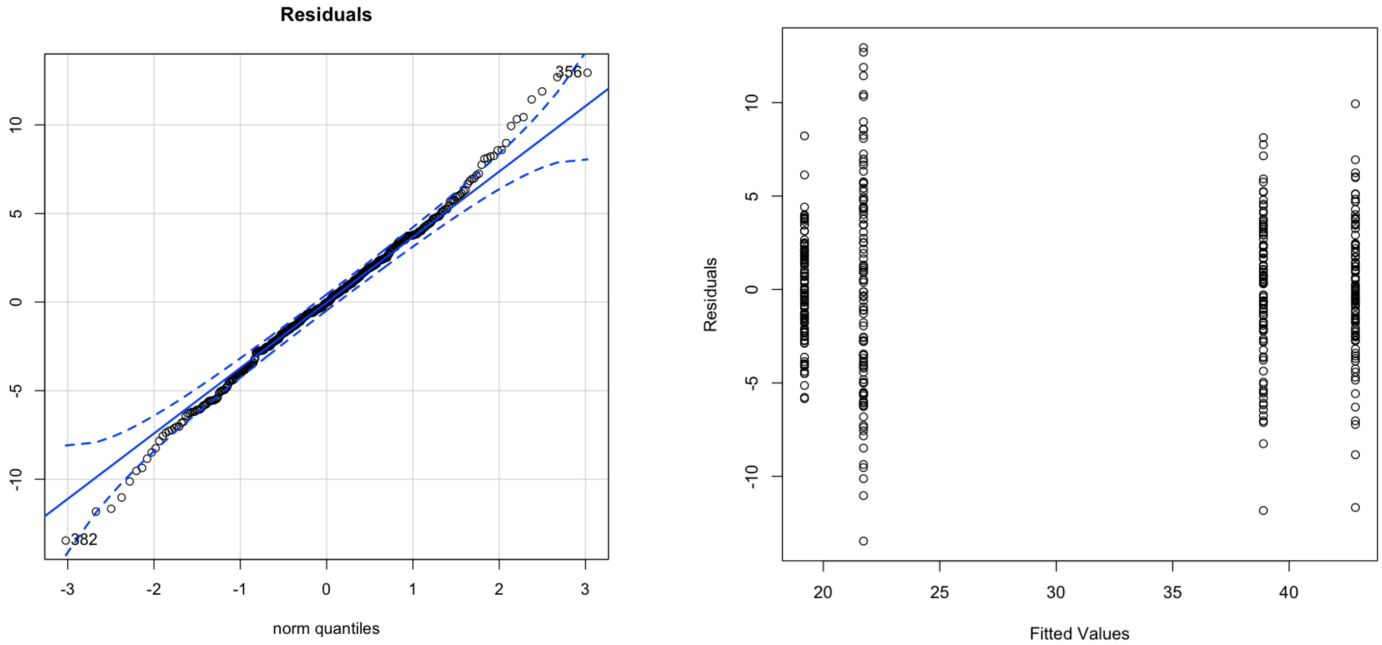


Figure 1: *Pilot: Boxplot, performance comparison of all four operators*



(a) *Pilot: Normality test with QQ-Plot*

(b) *Pilot: Homoscedasticity visualization*

Figure 2: *Pilot: Normality and Homoscedasticity check visualization*

3 Formulation of the experimental hypothesis

The ANOVA and paired T-Test were chosen since the data seem to be normal and these two tests are quite robust against violation of homoscedasticity. The desired parameters are $\alpha = 0.95$, $\beta = 0.85$, and $\delta = 0.25$.

The ANOVA test is used to detect difference in performance between the four recombination operators. The null hypothesis is that the mean of all operators are the same, while the alternative hypothesis is that at least one operator has mean not equal to the others.

$$\begin{cases} H_0 : \mu_i = \mu_j & \forall i, j \quad where \quad i \neq j \\ H_1 : \mu_i \neq \mu_j & \exists i, j \quad where \quad i \neq j \end{cases}$$

After detecting difference in performance, multiples unpaired two-samples T-test is performed to detect difference of performance between the operators. The null hypothesis is that the mean performance of one operators is greater or equal than the other, while the alternative hypothesis is that the mean of one operator is less than the other. In this study case, the best operator should have the lesser mean.

$$\begin{cases} H_0 : \mu_a \geq \mu_b \\ H_1 : \mu_a < \mu_b \end{cases}$$

4 Estimation of the effect size and the confidence intervals

Multiple T-tests cause type 1 error inflation. To prevent this, the Sidak correction was used. In this way, each of the six T-test comparison should have a confidence level of 0.0085.

5 Calculation of the Sample Size

Using the parameters discussed on the previous sections, ANOVA requires a sample size of only 3, while T-test requires a sample size of 12198. However, a sample size of 7000 was chosen, since there are limitations of time and computer power. This decision may cause type 2 error, where we fail to reject the null hypothesis when it is false.

6 Collection of the experiment data

These operators were compared against the sphere function with maximum number of function evaluation = 6000 and maximum number of iteration = 1000. Other operator specific parameters can be found in the experiment script on the github repository [1]. A total of n = 7000 data points were collected for each operator.

Figure 3 shows the boxplot of the performance of all four recombination operators. As can be seen, the data presents the same proprierties indentified in the pilot study with the exception of having many outliers.

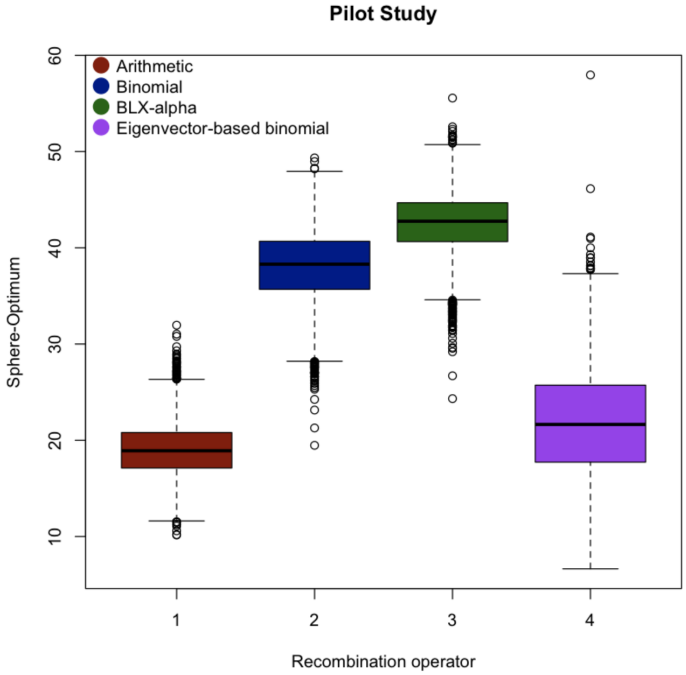


Figure 3: *Boxplot: performance comparison of all four operators*

7 Statistical Analysis and Hypothesis testing

ANOVA indicates that there is a significant difference of performance between the four operators (p-value < 2e-16). Table 1 shows the p-values for the pairwise T-test comparison of all operators. As can be seen, the operator Arithmetic is the best one. Next, Binomial is better than BLX-alpha, but it does not show statistical significant difference when compared to Eigenvector-based binomial. There is no clear difference between BLX-alpha and Eigenvector-based binomial.

	Arithmetic	**Binomial**	**BLX-alpha**
Binomial	<2.2e-16	/	/
BLX-alpha	<2.2e-16	<2.2e-16	/
Eigen-BB	<2.2e-16	1	1

Table 1: *P-Values for the pairwise T-test comparison*

8 Verification of the test assumptions

The data fails the assumptions of the parametric tests, despite the pilot study showing otherwise. Shapiro-test fails to reject the null hypothesis of non-normality with p-value = 1.19e-13. Figure 4-a shows the QQ-plot of the residuals. Contrary to the pilot study, there are strong indications that the data is heavy-tailed. As for homogeneity of variances, the results are similar to the pilot study due to the operator Eigenvector-based binomial having huge variance. The Fligner-Killeen test does not reject the null hypothesis with p-value ¡ 2.2e-16.

9 Further tests: non-parametric

Kruskal-Wallis test is an alternative for ANOVA when the data does not meet the parametric assumptions.Similarly, pairwise Wilcox is an option when T-test cannot be used. Kruskal-Wallis detected difference among the operators with p-value < 2.2e-16. Table 2 shows the p-values for the pairwise Wilcox test. As shown, it is exactly the same results as the parametric tests.

	Arithmetic	**Binomial**	**BLX-alpha**
Binomial	<2.2e-16	/	/
BLX-alpha	<2.2e-16	<2.2e-16	/
Eigen-BB	<2.2e-16	1	1

Table 2: *P-Values for the pairwise Wilcox test*

10 Conclusion

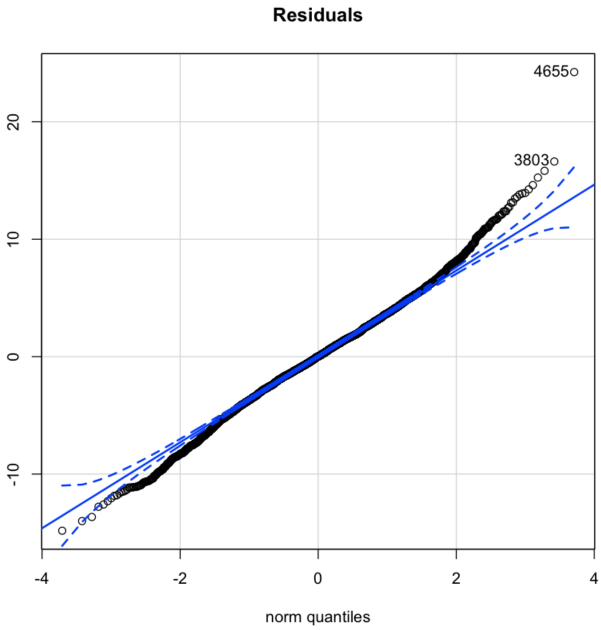
This study tried to verify if there is difference of performance between four different types of recombination operator for the Differential Algorithm. Although the parametric assumptions could not be met, the non-parametric and parametric tests presented exact the same results, indicating that the best operator for the sphere problem is the Arithmetic. Binomial is better than BLX-alpha, but it does not show statistical significant difference when compared to Eigenvector-based binomial. Also, there is no clear difference between BLX-alpha and Eigenvector-based binomial.

11 Discussions of possible limitations of the experiment, and suggestion for improvement

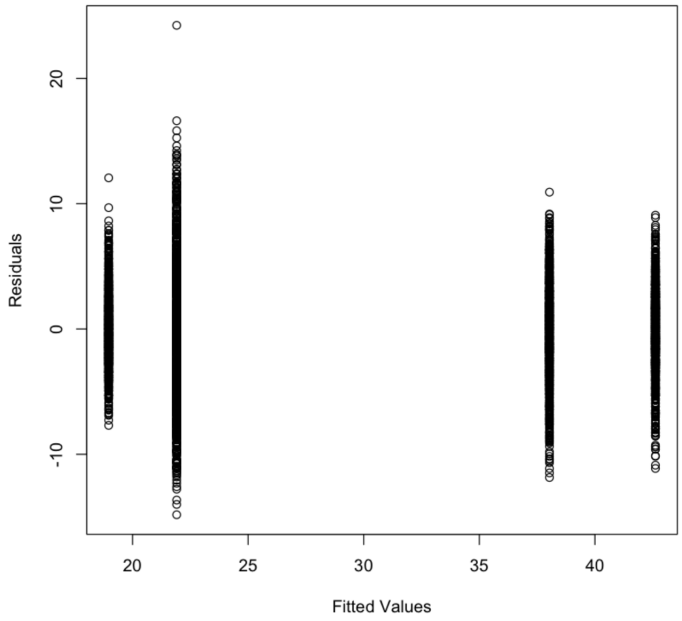
This experiment helped me understand better how to use the statistical tests and how to interpret better the QQ-plot and Boxplot. I am still have some questions: 1) Assuming that I did everything right, the alpha correction indicated a quite high sample size, was it expected? 2) How decide if it is ok to proceed with parametric tests when the assumptions have failed? 3) Had I chosen a considerable smaller sample size (n=10) for the pilot study, there would be good chances of the power test indicating also smaller sample size for the real experiment analysis. In this situation, we would not know that the data is heavy-tailed and not normal. How should we pick a good and feasible sample size for the pilot study?

References

[1] Pereira Junior, Jair. 2019. Github Repository: Experimental_Design_19_HW2 https://github.com/jair-pereira/Experimental_Design_19_HW3



(a) *Normality test with QQ-Plot*



(b) *Homoscedasticity visualization*

Figure 4: *Normality and Homoscedasticity check visualization*