J Heuristics (2009) 15: 617–644
DOI 10.1007/s10732-008-9080-4

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

Salvador García · Daniel Molina · Manuel Lozano · Francisco Herrera

Received: 24 October 2007 / Revised: 21 February 2008 / Accepted: 25 April 2008 /
Published online: 14 May 2008

Abstract In recent years, there has been a growing interest for the experimental analysis in the field of evolutionary algorithms. It is noticeable due to the existence of numerous papers which analyze and propose different types of problems, such as the basis for experimental comparisons of algorithms, proposals of different methodologies in comparison or proposals of use of different statistical techniques in algorithms' comparison.

In this paper, we focus our study on the use of statistical techniques in the analysis of evolutionary algorithms' behaviour over optimization problems. A study about the required conditions for statistical analysis of the results is presented by using some models of evolutionary algorithms for real-coding optimization. This study is conducted in two ways: single-problem analysis and multiple-problem analysis. The results obtained state that a parametric statistical analysis could not be appropriate specially when we deal with multiple-problem results. In multiple-problem analysis, we propose the use of non-parametric statistical tests given that they are less restrictive than parametric ones and they can be used over small size samples of results. As a case study, we analyze the published results for the algorithms presented in the

This work was supported by Project TIN2005-08386-C05-01.
S. García holds a FPU scholarship from Spanish Ministry of Education and Science.

S. García (✉) · M. Lozano · F. Herrera
Department of Computer Science and Artificial Intelligence, University of Granada, Granada 18071, Spain
e-mail: salvagl@decsai.ugr.es

M. Lozano
e-mail: lozano@decsai.ugr.es

F. Herrera
e-mail: herrera@decsai.ugr.es

D. Molina
Department of Computer Engineering, University of Cádiz, Cádiz, Spain
e-mail: daniel.molina@uca.es

CEC'2005 Special Session on Real Parameter Optimization by using non-parametric test procedures.

Keywords Statistical analysis of experiments · Evolutionary algorithms · Parametric tests · Non-parametric tests

1 Introduction

The "No free lunch" theorem (Wolpert and Macready 1997) demonstrates that it is not possible to find one algorithm being better in behaviour for any problem. On the other hand, we know that we can work with different degrees of knowledge about the problem which we expect to solve, and that it is not the same to work without knowledge about the problem (hypothesis of the "no free lunch" theorem) than to work with partial knowledge about the problem, knowledge that allows us to design algorithms with specific characteristics which can make them more suitable for the solution of the problem.

Once situated in this field, the partial knowledge of the problem and the necessity of having disposals of algorithms for its solution, the question about deciding when an algorithm is better than another one is suggested. In the case of the use of evolutionary algorithms, the latter may be done attending to the efficiency and/or effectiveness criteria. When theoretical results are not available in order to allow the comparison of the behaviour of the algorithms, we have to focus on the analysis of empirical results.

In the last years, there has been a growing interest in the analysis of experiments in the field of evolutionary algorithms. The work of Hooker is pioneer in this line and it shows an interesting study on what we must do and not do when we suggest the analysis of the behaviour of a metaheuristic about a problem (Hooker 1997).

In relation to the analysis of experiments, we can find three types of works: the study and design of test problems, the statistical analysis of experiments and experimental design.

- Several authors have focused their interest in the design of test problems which could be appropriate to do a comparative study among the algorithms. Focusing our attention to continuous optimization problems, which will be used in this paper, we can point out the pioneer papers of Whitley and co-authors for the design of complex test functions for continuous optimization (Whitley et al. 1995, 1996), and the recent works of Gallagher and Yuan (2006); Yuan and Gallagher (2003). In the same way, we can find papers that present test cases for different types of problems.
- Centred on the statistical analysis of the results, if we analyze the published papers in specialized journals, we find that the majority of the articles make a comparison of results based on average values of a set of executions over a concrete case. In proportion, a little set of works use statistical procedures in order to compare results, although their use is recently growing and it is being suggested as a need for many reviewers. When we find statistical studies, they are usually based on the average and variance by using parametric

tests (ANOVA, t-test, etc....) (Czarn et al. 2004; Ozcelik and Erzurumlu 2006; Rojas et al. 2002). Recently, non-parametric statistical procedures have been considered for being used in analysis of results (García et al. 2007; Moreno-Pérez et al. 2007). A similar situation can be found in the machine learning community (Demšar 2006).

- The experimental design consists of a set of techniques which comprise methodologies for adjusting the parameters of the algorithms depending on the settings used and results obtained (Bartz-Beielstein 2006; Kramer 2007). In our study, we are not interested in this topic; we assume that the algorithms in a comparison have obtained the best possible results, depending on an optimal adjustment of their parameters in each problem.

We are interested in the use of statistical techniques for the analysis of the behaviour of the evolutionary algorithms over optimization problems, analyzing the use of the parametric statistical tests and the non-parametric ones (Sheskin 2003; Zar 1999). We will analyze the required conditions for the usage of the parametric tests, and we will carry out an analysis of results by using non-parametric tests.

The study of this paper will be organized into two parts. The first one, we will denoted it by *single-problem analysis*, corresponds to the study of the required conditions of a safe use of parametric statistical procedures when comparing the algorithms over a single problem. The second one, denoted by *multiple-problem analysis*, will suppose the study of the same required conditions when considering a comparison of algorithms over more than one problems simultaneously.

The single-problem analysis is usually found in specialized literature (Bartz-Beielstein 2006; Ortiz-Boyer et al. 2007). Although the required conditions for using parametric statistics are usually not fulfilled, as we will see here, a parametric statistical study could obtain similar conclusions to a non-parametric one. However, in the multiple-problem analysis, due to the dissimilarities in the results obtained and the small size of the sample to be analyzed, a parametric test may reach erroneous conclusions. In recent papers, authors start using single-problem and multiple-problem analysis simultaneously (Ortiz-Boyer et al. 2007).

Non-parametric tests can be used for comparing algorithms whose results represent average values for each problem, in spite of the inexistence of relationships among them. Given that the non-parametric tests do not require explicit conditions for being conducted, it is recommendable that the sample of results is obtained following the same criterion, that is, computing the same aggregation (average, mode, etc.) over the same number of runs for each algorithm and problem. They are used for analyzing the results of the CEC'2005 Special Session on Real Parameter Optimization (Suganthan et al. 2005) over all the test problems, in which average results of the algorithms for each function are published. We will show significant statistical differences among the algorithms compared in the CEC'2005 Special Session on Real Parameter Optimization, supporting the conclusions obtained in this session.

In order to do that, the paper is organized as follows. In Sect. 2, we describe the setting of the CEC'2005 Special Session: algorithms, tests functions and parameters. Section 3 shows the study on the required conditions for safe use of parametric tests, considering single-problem and multiple-problem analysis. We analyze the published results of the CEC'2005 Special Session on Real Parameter Optimization by using

non-parametric tests in Sect. 4. Section 5 points out some considerations on the use of non-parametric tests. The conclusions of the paper are presented in Sect. 6. An introduction to statistics and a complete description of the non-parametric tests procedures are given in Appendix A. The published average results of the CEC'2005 Special Session are shown in Appendix B.

2 Preliminaries: settings of the CEC'2005 Special Session

In this section we will briefly describe the algorithms compared, the test functions, and the characteristics of the experimentation in the CEC'2005 Special Session.

2.1 Evolutionary algorithms

In this section we enumerate the eleven algorithms which were presented in the CEC'2005 Special Session. For more details on the description and parameters used for each one, please refer to the respective contributions. The algorithms are: *BLX-GL50* (García-Martínez and Lozano 2005), *BLX-MA* (Molina et al. 2005), *CoEVO* (Pošík 2005), *DE* (Rönkkönen et al. 2005), *DMS-L-PSO* (Liang and Suganthan 2005), *EDA* (Yuan and Gallagher 2005), *G-CMA-ES* (Auger and Hansen 2005a), *K-PCX* (Sinha et al. 2005), *L-CMA-ES* (Auger and Hansen 2005b), *L-SaDE* (Qin and Suganthan 2005), *SPC-PNX* (Ballester et al. 2005).

2.2 Test functions

In the following we present the set of test functions designed for the Special Session on Real Parameter Optimization organized in the 2005 IEEE Congress on Evolutionary Computation (CEC 2005) (Suganthan et al. 2005).

It is possible to consult in Suganthan et al. (2005) the complete description of the functions, furthermore in the link the source code is included. The set of test functions is composed of the following functions:

- 5 Unimodals functions
  - Sphere function displaced.
  - Schwefel's problem 1.2 displaced.
  - Elliptical function rotated widely conditioned.
  - Schwefel's problem 1.2 displaced with noise in the fitness.
  - Schwefel's problem 2.6 with global optimum in the frontier.
- 20 Multimodals functions
  - 7 basic functions
    - Rosenbrock function displaced.
    - Griewank function displaced and rotated without frontiers.
    - Ackley function displaced and rotated with the global optimum in the frontier.
    - Rastrigin function displaced.
    - Rastrigin function displaced and rotated.
    - Weierstrass function displaced and rotated.
    - Schwefel's problem 2.13.

– 2 expanded functions.
– 11 hybrid functions. Each one of them have been defined through compositions of 10 out of the 14 previous functions (different in each case).

All functions have been displaced in order to ensure that their optima can never be found in the centre of the search space. In two functions, in addition, the optima can not be found within the initialization range, and the domain of search is not limited (the optimum is out of the range of initialization).

2.3 Characteristics of the experimentation

The experiments were performed following the instructions indicated in the document associated to the competition. The main characteristics are:

- Each algorithm is run 25 times for each test function, and the average of error of the best individual of the population is computed.
- We will use the study with dimension $D = 10$ and the algorithms do 100000 evaluations of the fitness function.

  In the mentioned competition, experiments with dimension $D = 30$ and $D = 50$ have also been done.
- Each run stops either when the error obtained is less than 10^{-8}, or when the maximal number of evaluations is achieved.

3 Study of the required conditions for the safe use of parametric tests

In this section, we will describe and analyze the conditions that must be satisfied for the safe usage of parametric tests (Sect. 3.1). For doing it, we collect the overall set of results obtained by the algorithms *BLX-MA* and *BLX-GL50* in the 25 functions considering dimension $D = 10$. With them, we will firstly analyze the indicated conditions over the complete sample of results for each function, in a single-problem analysis (see Sect. 3.2). Finally, we will consider the average results for each function to composite a sample of results for each one of the two algorithms. With these two samples we will check again the required conditions for the safe use of parametric test in a multiple-problem scheme (see Sect. 3.3).

3.1 Conditions for the safe use of parametric tests

In Sheskin (2003), the distinction between parametric and non-parametric tests is based on the level of measure represented by the data which will be analyzed. In this way, a parametric test uses data composed by real values.

The latter does not imply that when we always dispose of this type of data, we should use a parametric test. There are other initial assumptions for a safe usage of parametric tests. The non fulfillment of these conditions might cause a statistical analysis to lose credibility.

In order to use the parametric tests, it is necessary to check the following conditions (Sheskin 2003; Zar 1999):

- Independence: In statistics, two events are independent when the fact that one occurs does not modify the probability of the other one occurring.
- Normality: An observation is normal when its behaviour follows a normal or Gauss distribution with a certain value of average μ and variance σ. A normality test applied over a sample can indicate the presence or absence of this condition in observed data. We will use three normality tests:
  - Kolmogorov-Smirnov: It compares the accumulated distribution of observed data with the accumulated distribution expected from a Gaussian distribution, obtaining the p-value based on both discrepancies.
  - Shapiro-Wilk: It analyzes the observed data to compute the level of symmetry and kurtosis (shape of the curve) in order to compute the difference with respect to a Gaussian distribution afterwards, obtaining the p-value from the sum of the squares of these discrepancies.
  - D'Agostino-Pearson: It first computes the skewness and kurtosis to quantify how far from Gaussian the distribution is in terms of asymmetry and shape. It then calculates how far each of these values differs from the value expected with a Gaussian distribution, and computes a single p-value from the sum of these discrepancies.
- Heteroscedasticity: This property indicates the existence of a violation of the hypothesis of equality of variances. Levene's test is used for checking whether or not k samples present this homogeneity of variances (homoscedasticity). When observed data does not fulfill the normality condition, this test's result is more reliable than Bartlett's test (Zar 1999), which checks the same property.

In our case, it is obvious the independence of the events given that they are independent runs of the algorithm with randomly generated initial seeds. In the following, we will carry out the normality analysis by using Kolmogorov-Smirnov, Shapiro-Wilk and D'Agostino-Pearson tests on single-problem and multiple-problem analysis, and heteroscedasticity analysis by means of Levene's test.

3.2 On the study of the required conditions over single-problem analysis

With the samples of results obtained from running 25 times the algorithms *BLX-GL50* and *BLX-MA* for each function, we can apply statistical tests for determining whether they check or not the normality and homoscedasticity properties. We have seen before that the independence condition is easily satisfied in this type of experiments. The number of runs may be low for carrying out statistical analysis, but it was a requirement in the CEC'2005 Special Session.

All the tests used in this section will obtain the p-value associated, which represents the dissimilarity of the sample of results with respect to the normal shape. Hence, a low p-value points out a non-normal distribution. In this study, we will consider a level of significance $\alpha = 0.05$, so a p-value greater than α indicates that the condition of normality is fulfilled. All the computations have been performed by the statistical software package SPSS.

Table 1 shows the results where the symbol "*" indicates that the normality is not satisfied and the p-value in brackets. Table 2 shows the results by applying the test

Table 1 Test of normality of Kolmogorov-Smirnov

	f1	f2	f3	f4	f5	f6	f7	f8	f9
BLX-GL50	(.20)	* (.04)	* (.00)	(.14)	* (.00)	* (.00)	* (.04)	(.20)	* (.00)
BLX-MA	* (.01)	* (.00)	* (.01)	* (.00)	* (.00)	(.16)	(.20)	* (.00)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
BLX-GL50	(.10)	(.20)	* (.00)	(.20)	(.20)	* (.00)	* (.00)	(.20)	* (.00)
BLX-MA	(.20)	* (.00)	* (.00)	(.20)	* (.02)	* (.00)	(.20)	(.20)	* (.00)
	f19	f20	f21	f22	f23	f24	f25		
BLX-GL50	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)		
BLX-MA	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.02)		

Table 2 Test of normality of Shapiro-Wilk

	f1	f2	f3	f4	f5	f6	f7	f8	f9
BLX-GL50	* (.03)	(.06)	* (.00)	* (.03)	* (.00)	* (.00)	* (.01)	(.23)	* (.00)
BLX-MA	* (.00)	* (.00)	* (.01)	* (.00)	* (.00)	(.05)	(.27)	* (.03)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
BLX-GL50	(.07)	(.25)	* (.00)	(.39)	(.41)	* (.00)	* (.00)	(.12)	* (.00)
BLX-MA	(.31)	* (.00)	* (.00)	(.56)	* (.01)	* (.00)	(.25)	(.72)	* (.00)
	f19	f20	f21	f22	f23	f24	f25		
BLX-GL50	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)		
BLX-MA	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.00)	* (.02)		

of normality of Shapiro-Wilk and Table 3 displays the results of D'Agostino-Pearson test.

In addition to this general study, we show the sample distribution in three cases, with the objective of illustrating representative cases in which the normality tests obtain different results.

From Fig. 1 to Fig. 3, different examples of graphical representations of histograms and Q-Q graphics are shown. A histogram represents a statistical variable by using bars, so that the area of each bar is proportional to the frequency of the represented values. A Q-Q graphic represents a confrontation between the quartiles from data observed and those from the normal distributions.

In Fig. 1 we can observe a general case in which the property of abnormality is clearly presented. On the contrary, Fig. 2 is the illustration of a sample whose distribution follows a normal shape, and the three normality tests employed verified this fact. Finally, Fig. 3 shows a special case where the similarity between both distributions, the sample of results and the normal one, is not confirmed by all normality

Table 3 Test of normality of D'Agostino-Pearson

	f1	f2	f3	f4	f5	f6	f7	f8	f9
BLX-GL50	(.10)	(.06)	* (.00)	(.24)	* (.00)	* (.00)	(.28)	(.21)	* (.00)
BLX-MA	* (.00)	* (.00)	(.22)	* (.00)	* (.00)	* (.00)	(.19)	(.12)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
BLX-GL50	(.17)	(.19)	* (.00)	(.79)	(.47)	* (.00)	* (.00)	(.07)	* (.03)
BLX-MA	(.89)	* (.00)	* (.03)	(.38)	(.16)	* (.00)	(.21)	(.54)	* (.04)
	f19	f20	f21	f22	f23	f24	f25		
BLX-GL50	(.05)	(.05)	(.06)	* (.01)	* (.00)	* (.00)	(.11)		
BLX-MA	* (.00)	* (.00)	(.25)	* (.00)	* (.00)	* (.00)	(.20)		

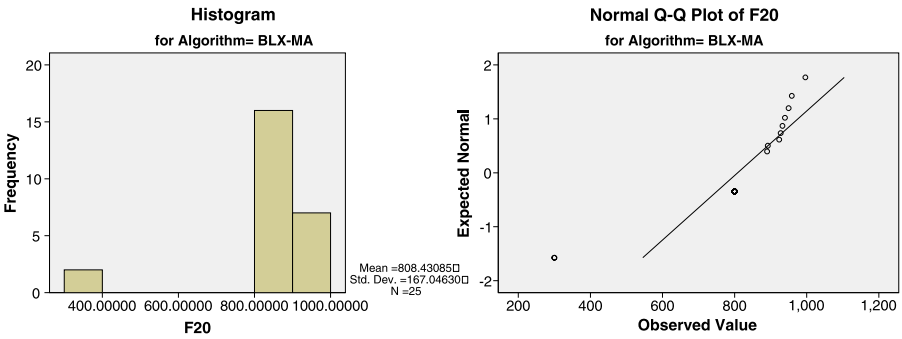

Fig. 1 Example of non-normal distribution: Function f20 and BLX-GL50 algorithm: Histogram and Q-Q Graphic

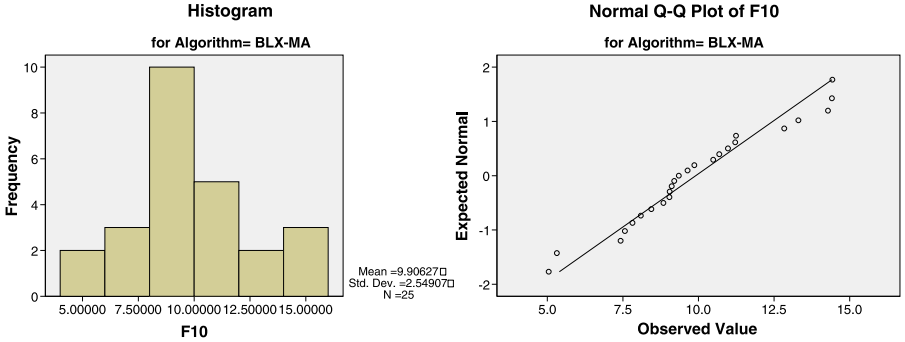

Fig. 2 Example of normal distribution: Function f10 and BLX-MA algorithm: Histogram and Q-Q Graphic

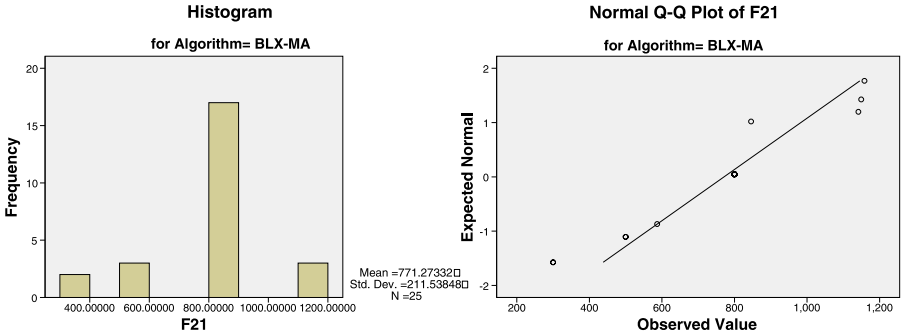

Fig. 3 Example of a special case: Function f21 and BLX-MA algorithm: Histogram and Q-Q Graphic

Table 4 Test of heteroscedasticity of Levene (based on means)

	f1	f2	f3	f4	f5	f6	f7	f8	f9
LEVENE	(.07)	(.07)	* (.00)	* (.04)	* (.00)	* (.00)	* (.00)	(.41)	* (.00)
	f10	f11	f12	f13	f14	f15	f16	f17	f18
LEVENE	(.99)	* (.00)	(.98)	(.18)	(.87)	* (.00)	* (.00)	(.24)	(.21)
	f19	f20	f21	f22	f23	f24	f25		
LEVENE	* (.01)	* (.00)	* (.01)	(.47)	(.28)	* (.00)	* (.00)		

tests. In this case, a normality test could work better than another, depending on types of data, number of ties or number of results collected. Due to this fact, we have employed three well-known normality tests for studying the normality condition. The choice of the most appropriate normality test depending on the problem is out of the scope of this paper.

With respect to the study of homoscedasticity property, Table 4 shows the results by applying Levene's test, where the symbol "*" indicates that the variances of the distributions of the different algorithms for a certain function are not homogeneities (we reject the null hypothesis at a level of significance $\alpha = 0.05$).

Clearly, in both cases, the non fulfillment of the normality and homoscedasticity conditions is perfectible. In most functions, the normality condition is not verified in a single-problem analysis. The homoscedasticity is also dependent of the number of algorithms studied, because it checks the relationship among the variances of all population samples. Even though in this case we only analyze this condition on results for two algorithms, the condition is also not fulfilled in many cases.

A researcher may think that the non fulfillment of these conditions is not crucial for obtaining adequate results. By using the same samples of results, we will show an example in which some results offered by a parametric test, the paired t-test, do not agree with the ones obtained through a non-parametric test, Wilcoxon's test. Table 5 presents the difference of average error rates, in each function, between the

Table 5 Difference of error rates and p-values for paired t-test and Wilcoxon test in single-problem analysis

Function	Difference	t-test	Wilcoxon
f1	0	–	–
f2	0	–	–
f3	−47129	0	0
f4	$-1.9 \cdot 10^{-8}$	0.281	0
f5	−0.0212	0.011	0
f6	−1.489618	0	0
f7	−0.1853	0	0
f8	0.2	0.686	0.716
f9	0.716	0	0
f10	−0.668086	0	0
f11	−2.223405	0.028	0.037
f12	332.7	0.802	0.51
f13	−0.024	0.058	0.058
f14	0.142023	0.827	0.882
f15	130	0.01	0.061
f16	−8.5	0	0
f17	−18	0	0
f18	−383	0	0
f19	−314	0	0.001
f20	−354	0	0
f21	−33	0.178	0.298
f22	88	0.545	0.074
f23	−288	0	0
f24	−24	0.043	0.046
f25	8	0.558	0.459

algorithms *BLX-GL50* and *BLX-MA* (if it is negative, the best performed algorithm is *BLX-GL50*), and the p-value obtained by the paired t-test and Wilcoxon test.

As we can see, the p-values obtained by paired t-test are very similar to the ones obtained by Wilcoxon test. However, in three cases, they are quite different. We enumerate them:

- In function f4, Wilcoxon test considers that both algorithms behave differently, whereas paired t-test does not. This example perfectly fits with a non-practical case. The difference of error rates is less than 10^{-7}, and in practical sense, this has no significant effect.
- In function f15, the situation is opposite to the previous one. The paired t-test obtains a significant difference in favour of *BLX-MA*. Is this result reliable? As the normality condition is not verified in the results of f15 (see Tables 1, 2, 3), the results obtained by Wilcoxon test are theoretically more reliable.
- Finally, in function f22, although Wilcoxon test obtains a p-value greater than the level of significance $\alpha = 0.05$, both p-values are again very different.

In 3 of the 25 functions, there are observable differences in the application of paired t-test and Wilcoxon test. Moreover, in these 3 functions, the required conditions for the safe usage of parametric statistics are not verified. In principle, we could suggest the usage of the non-parametric test of Wilcoxon in single-problem analysis. This is one alternative, but there exist other ways for ensuring that the results obtained are valid for parametric statistical analysis.

- Obtaining new results is not very difficult in single-problem analysis. We only have to run the algorithms again to get larger samples of results. The Central Limit Theorem confirms that the sum of many identically distributed random variables tends to a normal distribution. Nevertheless, the number of runs carried out must not be very high, because any statistical test has a negative effect size. If the sample of results is too large, a statistical test could detect insignificant differences as significant.

  For controlling the size effect, we can use the Cohen's index d'

$$d' = \frac{t}{\sqrt{n}}$$

  where t is the t-test statistics and n is the number of results collected. If d' is near to 0.5, then the differences are significant. A value of d' lower than 0.25 indicates insignificant differences and the statistical analysis may not be taken into account.
- The application of transformations for obtaining normal distributions, such as logarithm, square root, reciprocal and power transformations (Patel and Read 1982).
- In some situations, skip outliers, but this technique must be used with great care.

These alternatives could solve the normality condition, but the homoscedasticity condition may result difficult to solve. Some parametric tests, such as ANOVA, are very influenced by the homoscedasticity condition.

3.3 On the study of the required conditions over multiple-problem analysis

When tackling a multiple-problem analysis, the data to be used is an aggregation of results obtained from individual algorithms' runs. In this aggregation, there must be only a result representing a problem or function. This result could be obtained through averaging results for all runs or something similar, but the procedure followed must be the same for each function; i.e., in this paper we have used the average of the 25 runs of an algorithm in each function. The size of the sample of results to be analyzed, for each algorithm, is equal to the number of problems. In this way, a multiple-problem analysis allows us to compare two or more algorithms over a set of problems simultaneously.

We can use the results published in the CEC'2005 Special Session to perform a multiple-problem analysis. Indeed, we will follow the same procedure as the previous subsection. We will analyze the required conditions for the safe usage of parametric tests over the sample of results obtained by averaging the error rate on each function.

Table 6 shows the p-values of the normality tests over the sample results obtained by *BLX-GL50* and *BLX-MA*. Figures 4 and 5 represent the histograms and Q-Q plots for such samples.

Obviously, the normality condition is not satisfied because the sample of results is composed by 25 average error rates computed in 25 different problems. We compare the behaviour of the two algorithms by means of pairwise statistical tests:

- The p-value obtained with a paired t-test is $p = 0.318$. The paired t-test does not consider the existence of difference in performance between the algorithms.
- The p-value obtained with Wilcoxon test is $p = 0.089$. The Wilcoxon t-test does neither consider the existence of difference in performance between the algorithms, but it considerably reduces the minimal level of significance for detecting differences. If the level of significance considered were $\alpha = 0.10$, Wilcoxon's test would confirm that *BLX-GL50* is better than *BLX-MA*.

Average results for these two algorithms indicate this behaviour, *BLX-GL50* usually performs better than *BLX-MA* (see Table 13 in Appendix B), but a paired t-test

Table 6 Normality tests over multiple-problem analysis

Algorithm	Kolmogorov-Smirnov	Shapiro-Wilk	D'Agostino-Pearson
BLX-GL50	* (.00)	* (.00)	(.10)
BLX-MA	* (.00)	* (.00)	* (.00)

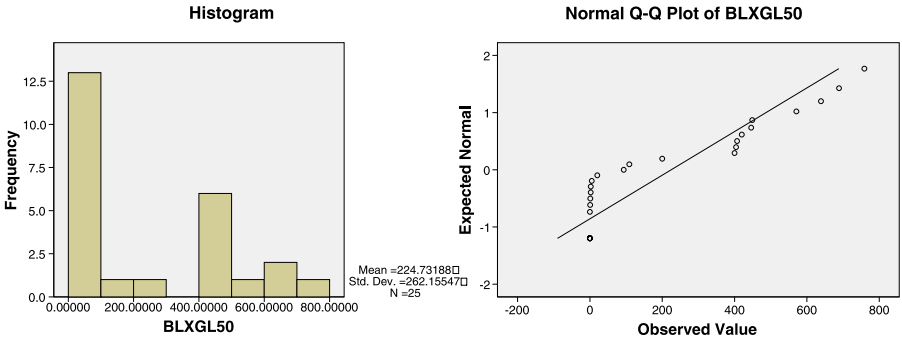

Fig. 4 BLX-GL50 algorithm: Histogram and Q-Q Graphic

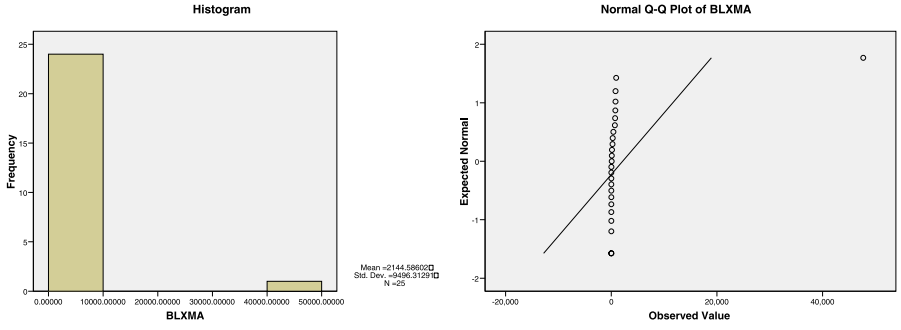

Fig. 5 BLX-MA algorithm: Histogram and Q-Q Graphic

cannot appreciate this fact. In multiple-problem analysis it is not possible to enlarge the sample of results, unless new functions/problems were added. Applying transformations or skipping outliers cannot be used either, because we would be changing results for certain problems and not for other problems.

These facts may induce us to using non-parametric statistics for analyzing the results in multiple-problems. Non-parametric statistics do not need prior assumptions related to the sample of data for being analyzed and, in the example shown in this section, we have seen that they could obtain reliable results.

4 A case study: on the use of non-parametric statistics for comparing the results of the CEC'2005 Special Session in Real Parameter Optimization

In this section, we study the results obtained in the CEC'2005 Special Session in Real Parameter Optimization as a case study on the use of the non-parametric tests. As we have mentioned, we will focus on the dimension $D = 10$.

We will divide the set of functions into two subgroups, according to the suggestion given in Hansen (2005) about their degrees of difficulty.

- The first group is composed by the unimodal functions (from f1 to f5), in which all participant algorithms in the CEC'2005 competition normally achieve the optimum, and the multimodal functions (from f6 to f14), in which at least one run of a participant algorithm achieves the optimum.
- The second group contains the remaining functions, from the function f15 to f25. In these functions, no participant algorithm has achieved the optimum.

This division is carried out with the objective of showing the differences in the statistical analysis considering distinct numbers of functions, which is an essential factor that influences over the study. It also allows us to compare the behaviour of the algorithms when they tackle the most complicated functions. Indeed, we could also study the group of functions f1–f14, but we do not include it in order not to enlarge the content of the paper. Hence, the results offered by the algorithms that take part in the CEC'2005 Special Session are analyzed independently for all functions (from f1 to f25) and the difficult functions (from f15 to f25).

As we have done before, we have considered using, as performance measure, the error rate obtained for each algorithm. This case corresponds to a multiple-problem analysis, so the employment of non-parametric statistical tests is preferable to a parametric one, as we have seen in the previous section. Table 13 in Appendix B summarizes the official results obtained in the competition organized by functions and algorithms.

Values included in Table 13 allow us to carry out a rigorous statistical study in order to check whether the results of the algorithms are rather significant for considering them different in terms of quality on approximation of continuous functions. Our study will be focused on the algorithm that had the lowest average error rate in the comparison, *G-CMA-ES* (Hansen 2005). We will study the behaviour of this algorithm with respect to the remaining ones, and we will determine if the results it offers are better than the ones offered by the rest of algorithms, computing the p-values on each comparison.

Table 7 Results of the Friedman and Iman-Davenport tests ($\alpha = 0.05$)

	Friedman value	Value in χ^2	p-value	Iman-Davenport value	Value in F_F	p-value
f15–f25	**26.942**	18.307	0.0027	**3.244**	1.930	0.0011
All	**41.985**	18.307	<0.0001	**4.844**	1.875	<0.0001

Table 8 Rankings obtained through Friedman's test and critical difference of Bonferroni-Dunn's procedure

Algorithm	Ranking (f15–f25)	Ranking (f1–f25)
BLX-GL50	5.227	5.3
BLX-MA	7.681	7.14
CoEVO	9.000	6.44
DE	4.955	5.66
DMS-L-PSO	5.409	5.02
EDA	6.318	6.74
G-CMA-ES	3.045	3.34
K-PCX	7.545	6.8
L-CMA-ES	6.545	6.22
L-SaDE	4.956	4.92
SPC-PNX	5.318	6.42
Crit. Diff. $\alpha = 0.05$	**3.970**	**2.633**
Crit. Diff. $\alpha = 0.10$	**3.643**	**2.417**

Table 7 shows the result of applying Friedman's and Iman-Davenport's tests in order to see whether there are global differences in the results. Given that the p-values of Friedman and Iman-Davenport are lower than the level of significance considered $\alpha = 0.05$, there are significant differences among the observed results in the functions of the first and second group. Attending to these results, a post-hoc statistical analysis could help us to detect concrete differences among algorithms.

First of all, we will employ Bonferroni-Dunn's test to detect significant differences for the control algorithm *G-CMA-ES*. Table 8 summarizes the ranking obtained by Friedman's test and the critical difference of Bonferroni-Dunn's procedure. Figures 6 and 7 display graphical representations (including the rankings obtained for each algorithm) for the two groups of functions. In a Bonferroni-Dunn's graphic the difference among rankings obtained for each algorithm is illustrated. In them, we can draw a horizontal cut line which represents the threshold for the best performing algorithm, that one with the lowest ranking bar, in order to consider it better than other algorithm. A cut line is drawn for each level of significance considered in the study at height equal to the sum of the ranking of the control algorithm and the corresponding Critical Difference computed by the Bonferroni-Dunn method (see Appendix A.3). Those bars which exceed this line are the associated to an algorithm with worse performance than the control algorithm.

The application of Bonferroni-Dunn's test informs us of the following significant differences with *G-CMA-ES* as control algorithm:

- f15–f25: *G-CMA-ES* is better than *CoEVO* and *BLX-MA* and *K-PCX* with $\alpha = 0.05$ and $\alpha = 0.10$ (3/10 algorithms).
- f1–f25: It outperforms *CoEVO*, *BLX-MA*, *K-PCX*, *EDA*, *SPC-PNX* and *L-CMA-ES* with $\alpha = 0.05$ and $\alpha = 0.10$ (6/10 algorithms). Although *G-CMA-ES* obtains the lowest error and ranking rates, Bonferroni-Dunn's test is not able to distinguish it as better than all the remaining algorithms.

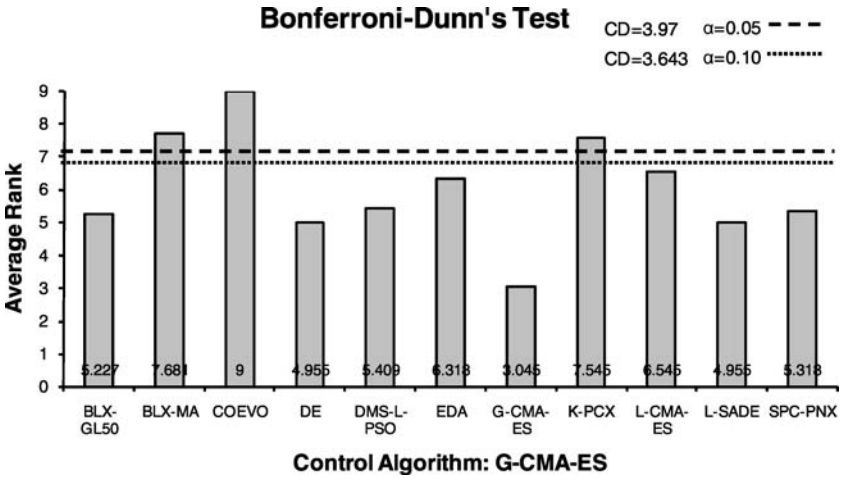

Fig. 6 Bonferroni-Dunn's graphic corresponding to the results for f15–f25

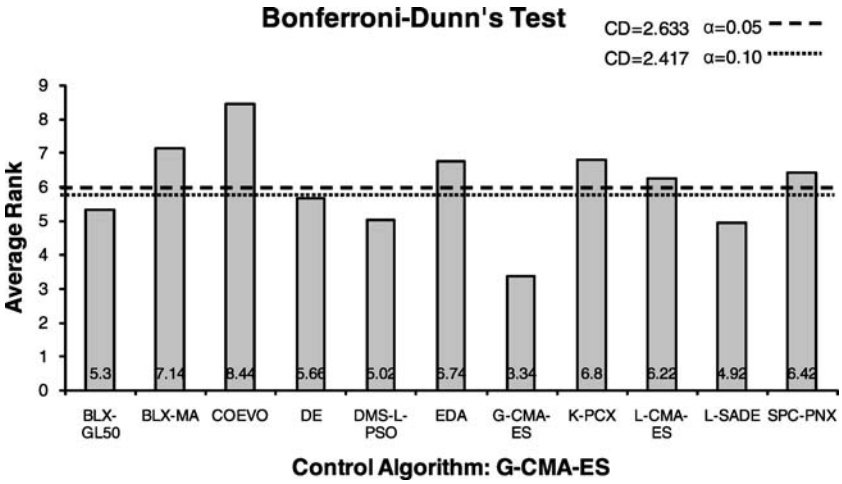

Fig. 7 Bonferroni-Dunn's graphic corresponding to the results for f1–f25

Table 9 p-values on functions f15–f25 (G-CMA-ES is the control algorithm)

G-CMA-ES vs.	z	Unadjusted p	Bonferroni-Dunn p	Holm p	Hochberg p
CoEVO	4.21050	$2.54807 \cdot 10^{-5}$	$2.54807 \cdot 10^{-4}$	$2.54807 \cdot 10^{-4}$	$2.54807 \cdot 10^{-4}$
BLX-MA	3.27840	0.00104	0.0104	0.00936	0.00936
k-PCX	3.18198	0.00146	0.0146	0.01168	0.01168
L-CMA-ES	2.47487	0.01333	0.1333	0.09331	0.09331
EDA	2.31417	0.02066	0.2066	0.12396	0.12396
DMS-L-PSO	1.67134	0.09465	0.9465	0.47325	0.17704
SPC-NPX	1.60706	0.10804	1.0	0.47325	0.17704
BLX-GL50	1.54278	0.12288	1.0	0.47325	0.17704
DE	1.34993	0.17704	1.0	0.47325	0.17704
L-SaDE	1.34993	0.17704	1.0	0.47325	0.17704

Table 10 p-values on functions f1–f25 (G-CMA-ES is the control algorithm)

G-CMA-ES vs.	z	Unadjusted p	Bonferroni-Dunn p	Holm p	Hochberg p
CoEVO	5.43662	$5.43013 \cdot 10^{-8}$	$5.43013 \cdot 10^{-7}$	$5.43013 \cdot 10^{-7}$	$5.43013 \cdot 10^{-7}$
BLX-MA	4.05081	$5.10399 \cdot 10^{-5}$	$5.10399 \cdot 10^{-4}$	$4.59359 \cdot 10^{-4}$	$4.59359 \cdot 10^{-4}$
K-PCX	3.68837	$2.25693 \cdot 10^{-4}$	0.002257	0.001806	0.001806
EDA	3.62441	$2.89619 \cdot 10^{-4}$	0.0028961	0.002027	0.002027
SPC-PNX	3.28329	0.00103	0.0103	0.00618	0.00618
L-CMA-ES	3.07009	0.00214	0.0214	0.0107	0.0107
DE	2.47313	0.01339	0.1339	0.05356	0.05356
BLX-GL50	2.08947	0.03667	0.3667	0.11	0.09213
DMS-L-PSO	1.79089	0.07331	0.7331	0.14662	0.09213
L-SaDE	1.68429	0.09213	0.9213	0.14662	0.09213

In the same way as the previous section, we will apply more powerful procedures, such as Holm's and Hochberg's (they are described in Appendix A.3), for comparing the control algorithm with the rest of the algorithms. The results are shown by computing p-values for each comparison. Tables 9 and 10 show the p-value obtained for Bonferroni-Dunn's, Holm's and Hochberg's procedures considering both groups of functions. The procedure used to compute the p-values is explained in Appendix A.3.

Holm's and Hochberg's procedures allow us to point out the following differences, considering *G-CMA-ES* as control algorithm:

- f15–f25: *G-CMA-ES* is better than *CoEVO*, *BLX-MA* and *K-PCX* with $\alpha = 0.05$ (3/10 algorithms) and is better than *L-CMA-ES* with $\alpha = 0.10$ (4/10 algorithms). Here, Holm's and Hochberg's procedures coincide and they reject an extra hypothesis considering $\alpha = 0.10$, with regards to Bonferroni-Dunn's.
- f1–f25: Based on Holm's procedure, it outperforms *CoEVO*, *BLX-MA*, *K-PCX*, *EDA*, *SPC-PNX* and *L-CMA-ES* with $\alpha = 0.05$ (6/10 algorithms) and it also outperforms *DE* with $\alpha = 0.10$ (7/10 algorithms). It rejects equal number of hypotheses

as Bonferroni-Dunn does by considering $\alpha = 0.05$. It also rejects an extra hypothesis than Bonferroni-Dunn when $\alpha = 0.10$.

- Hochberg's procedure behaves the same as Holm's when we establish $\alpha = 0.05$. However, with a $\alpha = 0.10$, it obtains a different result. All the p-values in the comparison are lower than 0.10, so all the hypotheses associated with them are rejected (10/10 algorithms). In fact, Hochberg's procedure confirms that *G-CMA-ES* is the best algorithm in the competition considering all functions on the whole.

In the following, we present a study in which the *G-CMA-ES* algorithm will be compared with the rest of them by means of pairwise comparisons. In this study we will use the Wilcoxon test (see Appendix A.2).

Until now, we have used procedures for performing multiple comparisons in order to check the behaviour of the algorithms. Attending to Hochberg's procedure results, this process could not be necessary, but we include it for stressing the differences between using multiple comparisons procedures instead of pairwise comparisons. Tables 11 and 12 summarize the results of applying Wilcoxon test. They display the sum of rankings obtained in each comparison and the p-value associated.

Table 11 Wilcoxon test considering functions f15–f25

G-CMA-ES vs.	R^+	R^-	p-value
BLX-GL50	62.5	3.5	0.009
BLX-MA	60.0	6.0	0.016
CoEVO	60.0	6.0	0.016
DE	56.5	9.5	0.028
DMS-L-PSO	47.0	19.0	0.213
EDA	60.5	5.5	0.013
K-PCX	60.0	6.0	0.016
L-CMA-ES	58.0	8.0	0.026
L-SaDE	47.5	18.5	0.203
SPC-PNX	63.5	2.5	0.007

Table 12 Wilcoxon test considering functions f1–f25

G-CMA-ES vs.	R^+	R^-	p-value
BLX-GL50	289.5	35.5	0.001
BLX-MA	295.5	29.5	0.001
CoEVO	301.0	24.0	0.000
DE	262.5	62.5	0.009
DMS-L-PSO	199.0	126.0	0.357
EDA	284.5	40.5	0.001
K-PCX	269.0	56.0	0.004
L-CMA-ES	273.0	52.0	0.003
L-SaDE	209.0	116.0	0.259
SPC-PNX	305.5	19.5	0.000

Wilcoxon's test performs individual comparisons between two algorithms (pairwise comparisons). The p-value in a pairwise comparison is independent from another one. If we try to extract a conclusion involving more than one pairwise comparison in a Wilcoxon's analysis, we will obtain an accumulated error coming from the combination of pairwise comparisons. In statistical terms, we are losing the control on the Family Wise Error Rate (FWER), defined as the probability of making one or more false discoveries among all the hypotheses when performing multiple pairwise tests. The true statistical significance for combining pairwise comparisons is given by:

$$\begin{aligned} p &= P(\textit{Reject } H_0 | H_0 \textit{ true}) \\ &= 1 - P(\textit{Accept } H_0 | H_0 \textit{ true}) \\ &= 1 - P(\textit{Accept } A_k = A_i, i = 1, \ldots, k-1 | H_0 \textit{ true}) \\ &= 1 - \prod_{i=1}^{k-1} P(\textit{Accept\_}A_k = A_i | H_0 \textit{ true}) \\ &= 1 - \prod_{i=1}^{k-1} [1 - P(\textit{Reject } A_k = A_i | H_0 \textit{ true})] \\ &= 1 - \prod_{i=1}^{k-1} (1 - p_{H_i}) \end{aligned} \tag{1}$$

Observing Table 11, the statement: "The *G-CMA-ES* algorithm outperforms the *BLX-GL50*, *BLX-MA*, *CoEVO*, *DE*, *EDA*, *K-PCX*, *L-CMA-ES* and *SPC-PNX* algorithms with a level of significance $\alpha = 0.05$" could not be correct until we cannot check controlling the FWER. The *G-CMA-ES* algorithm really outperforms these eight algorithms considering independent pairwise comparisons due to the fact that the p-values are below $\alpha = 0.05$. On the other hand, note that two algorithms were not included. If we include them within the multiple comparison, the p-value obtained is $p = 0.4505$ in f15–f25 group and $p = 0.5325$ considering all functions. In such cases, it is not possible to declare that "*G-CMA-ES* algorithm obtains a significantly better performance than the remaining algorithms", due to the fact that the p-values achieved are too high.

From expression (1), and Tables 11 and 12, we can deduce that *G-CMA-ES* is better than the eight algorithms enumerated before with a p-value of

$$\begin{aligned} p = 1 - ((1 - 0.009) \cdot (1 - 0.016) \cdot (1 - 0.016) \cdot (1 - 0.028) \cdot (1 - 0.013) \\ \cdot (1 - 0.016) \cdot (1 - 0.026) \cdot (1 - 0.007)) = 0.123906 \end{aligned}$$

for the group of functions f15–f25 and

$$\begin{aligned} p = 1 - ((1 - 0.001) \cdot (1 - 0.001) \cdot (1 - 0.000) \cdot (1 - 0.009) \cdot (1 - 0.001) \\ \cdot (1 - 0.004) \cdot (1 - 0.003) \cdot (1 - 0.000)) = 0.018874 \end{aligned}$$

considering all functions. Hence, the previous statement has been definitively confirmed only when considering all functions in the comparison.

The procedures designed for performing multiple comparisons control the FWER in their definition. By using the example considered in this section, in which we have used the *G-CMA-ES* algorithm as control, we can easily reflect the relationship among the power of all the testing procedures used. In increasing order of power and considering all functions in the study, the procedures can be order in the following way: Bonferroni-Dunn ($p = 0.9213$), Wilcoxon's test (when it is used in multiple comparisons) ($p = 0.5325$), Holm ($p = 0.1466$) and Hochberg ($p = 0.0921$).

Finally, we must point out that the statistical procedures used here indicate that the best algorithm is *G-CMA-ES*. Although in Hansen (2005), the categorization of the functions depending on their degree of difficulty is different than the used in this paper (we have joined the unimodal and soluble multimodal functions in one group), the *G-CMA-ES* algorithm has been stressed as the algorithm with best behaviour considering error rate. Therefore and to sum up, in this paper the conclusions drawn in Hansen (2005) have been statistically supported.

5 Some considerations on the use of non-parametric tests

Taking into consideration all the results, tables and figures on the application of the non-parametric tests shown in this paper, we can suggest some aspects and details about the use of non-parametric statistical techniques:

- A multiple comparison of various algorithms must be carried out first by using a statistical method for testing the differences among the related samples means, that is, the results obtained by each algorithm. Once this test rejects the hypothesis of equivalence of means, the detection of the concrete differences among the algorithms can be done with the application of post-hoc statistical procedures, which are methods used for comparing a control algorithm with two or more algorithms.
- Holm's procedure can always be considered better than Bonferroni-Dunn's one, because it appropriately controls the FWER and it is more powerful than the Bonferroni-Dunn's. We strongly recommend the use of Holm's method in a rigorous comparison. Nevertheless, the results offered by the Bonferroni-Dunn's test are suitable to be visualized in graphical representations.
- Hochberg's procedure is more powerful than Holm's. The differences reported between it and Holm's procedure are in practice rather small, but in this paper, we have shown a case in which Hochberg's method obtains lower p-values than Holm's (see Table 10). We recommend the use of this test together with Holm's method.
- Although Wilcoxon's test and the remaining post-hoc tests for multiple comparisons belong to the non-parametric statistical tests, they operate in a different way. The main difference lies in the computation of the ranking. Wilcoxon's test computes a ranking based on differences between functions independently, whereas Friedman and derivative procedures compute the ranking between algorithms.
- In relation to the sample size (number of functions when performing Wilcoxon's or Friedman's tests in multiple-problem analysis), there are two main aspects to be

determined. Firstly, the minimum sample considered acceptable for each test needs to be stipulated. There is no established agreement about this specification. Statisticians have studied the minimum sample size when a certain power of the statistical test is expected (Noether 1987; Morse 1999). In our case, the employment of a, as large as possible, sample size is preferable, because the power of the statistical tests (defined as the probability that the test will reject a false null hypothesis) will increase. Moreover, in a multiple-problem analysis, the increasing of the sample size depends on the availability of new functions (which should be well-known in real-parameter optimization field). Secondly, we have to study how the results are expected to vary if there was a larger sample size available. In all statistical tests used for comparing two or more samples, the increasing of the sample size benefits the power of the test. In the following items, we will state that Wilcoxon's test is less influenced by this factor than Friedman's test. Finally, as a rule of thumb, the number of functions (N) in a study should be $N = a \cdot k$, where k is the number of algorithms to be compared and $a \geq 2$.

- Taking into account the previous observation and knowing the operations performed by the non-parametric tests, we can deduce that Wilcoxon's test is influenced by the number of functions used. On the other hand, both the number of algorithms and functions are crucial when we refer to the multiple comparisons tests (such as Friedman's test), given that all the critical values depend on the value of N (see expressions in Appendix A.3). However, the increasing/decreasing of the number of functions rarely affects in the computation of the ranking. In these procedures, the number of functions used is an important factor to be considered when we want to control the FWER.
- An appropriate number of algorithms in contrast with an appropriate number of functions are needed to be used in order to employ each type of test. The number of algorithms used in multiple comparisons procedures must be lower than the number of functions. In the study of the CEC'2005 Special Session, we can appreciate the effect of the number of functions used whereas the number of algorithms stays constant. See, for instance, the p-value obtained when considering the f15–f25 group and all functions. In the latter case, p-values obtained are always lower than in the first one, for each testing procedure. In general, p-values are lower agreeing with the increasing of the number of functions used in multiple comparison procedures; therefore, the differences among the algorithms are more detectable.
- The previous statement may not be true in Wilcoxon's test. The influence of the number of functions used is more noticeable in multiple comparisons procedures than in Wilcoxon's test. For example, the final p-value computed for Wilcoxon's test in group f15–f25 is lower than in the group f1–f25 (see previous section).

6 Conclusions

In this paper we have studied the use of statistical techniques in the analysis of the behaviour of evolutionary algorithms in optimization problems, analyzing the use of the parametric and non-parametric statistical tests.

We have distinguished two types of analysis. The first one, called single-problem analysis, is that in which the results are analyzed for each function/problem independently. The second one, called multiple-problem analysis, is that in which the results are analyzed by considering all the problems studied simultaneously.

In single-problem analysis, we have seen that the required conditions for a safe usage of parametric statistics are usually not satisfied. Nevertheless, the results obtained are quite similar between a parametric and non-parametric analysis. Also, there are procedures for transforming or adapting sample results for being used by parametric statistical tests.

We encourage the use of non-parametric tests when we want to analyze results obtained by evolutionary algorithms for continuous optimization problems in multiple-problem analysis, due to the fact that the initial conditions that guarantee the reliability of the parametric tests are not satisfied. In this case, the results come from different problems and it is not possible to analyze the results by means of parametric statistics.

With respect to the use of non-parametric tests, we have shown how to use Friedman, Iman-Davenport, Bonferroni-Dunn, Holm, Hochberg, and Wilcoxon's tests; which on the whole, are a good tool for the analysis of the algorithms. We have employed these procedures to carry out a comparison on the CEC'2005 Special Session on Real Parameter Optimization by using the results published for each algorithm.

Acknowledgements The authors are very grateful to the anonymous reviewers for their valuable suggestions and comments to improve the quality of this paper.

Appendix A: introduction to inferential statistical tests

This section is dedicated to introduce the necessary issues to understand the statistical terms used in this paper. Moreover, a description of the non-parametric tests is given in order to use them in further research. In order to distinguish a non-parametric test from a parametric one, we must check the type of data used by the test. A non-parametric test is that which uses nominal or ordinal data. This fact does not force it to be used only for these types of data. It is possible to transform the data from real values to ranking based data. In such way, a non-parametric test can be applied over classical data of parametric test when they do not verify the required conditions imposed by the test. As a general rule, a non-parametric test is less restrictive than a parametric one, although it is less robust than a parametric when data are well conditioned.

A.1 Hypothesis testing and p-values

In inferential statistics, sample data are primarily employed in two ways to draw inferences about one or more populations. One of them is the hypothesis testing.

The most basic concept in hypothesis testing is a hypothesis. It can be defined as a prediction about a single population or about the relationship between two or more

populations. Hypothesis testing is a procedure in which sample data are employed to evaluate a hypothesis. There is a distinction between research hypothesis and statistical hypothesis. The first is a general statement of what a researcher predicts. In order to evaluate a research hypothesis, it is restated within the framework of two statistical hypotheses. They are the null hypothesis, represented by the notation H_0, and the alternative hypothesis, represented by the notation H_1.

The null hypothesis is a statement of no effect or no difference. Since the statement of the research hypothesis generally predicts the presence of a difference with respect to whatever is being studied, the null hypothesis will generally be a hypothesis that the researcher expects to be rejected. The alternative hypothesis represents a statistical statement indicating the presence of an effect or a difference. In this case, the researcher generally expects the alternative hypothesis to be supported.

An alternative hypothesis can be nondirectional (two-tailed hypothesis) and directional (one-tailed hypothesis). The first type does not make a prediction in a specific direction; i.e. $H_1 : \mu \neq 100$. The latter implies a choice of one of the following directional alternative hypothesis; i.e. $H_1 : \mu > 100$ or $H_1 : \mu < 100$.

Upon collecting the data for a study, the next step in the hypothesis testing procedure is to evaluate the data through use of the appropriate inferential statistical test. An inferential statistical test yields a test statistic. The latter value is interpreted by employing special tables that contain information with regard to the expected distribution of the test statistic. Such tables contain extreme values of the test statistic (referred to as critical values) that are highly unlikely to occur if the null hypothesis is true. Such tables allow a researcher to determine whether or not the results of a study is statistically significant.

The conventional hypothesis testing model employed in inferential statistics assumes that prior to conducting a study, a researcher stipulates whether a directional or nondirectional alternative hypothesis is employed, as well as at what level of significance is represented the null hypothesis to be evaluated. The probability value which identifies the level of significance is represented by α.

When one employs the term significance in the context of scientific research, it is instructive to make a distinction between statistical significance and practical significance. Statistical significance only implies that the outcome of a study is highly unlikely to have occurred as a result of chance, but it does no necessarily suggest that any difference or effect detected in a set of data is of any practical value. For example, no-one would normally care if algorithm A solves the sphere function to within 10^{-10} of error of the global optimum and algorithm B solves it within 10^{-15}. Between them, statistical significance could be found, but in practical sense, this difference is not significant.

Instead of stipulating a priori a level of significance α, one could calculate the smallest level of significance that results in the rejection of the null hypothesis. This is the definition of p-value, which is an useful and interesting datum for many consumers of statistical analysis. A p-value provides information about whether a statistical hypothesis test is significant or not, and it also indicates something about "how significant" the result is: The smaller the p-value, the stronger the evidence against the null hypothesis. Most important, it does this without committing to a particular level of significance.

The most common way for obtaining the p-value associated to a hypothesis is by means of normal approximations, that is, once computed the statistic associated to a statistical test or procedure, we can use a specific expression or algorithm for obtaining a z value, which corresponds to a normal distribution statistics. Then, by using normal distribution tables, we could obtain the p-value associated with z.

A.2 The Wilcoxon matched-pairs signed-ranks test

Wilcoxon's test is used for answering this question: do two samples represent two different populations? It is a non-parametric procedure employed in a hypothesis testing situation involving a design with two samples. It is the analogous of the paired t-test in non-parametrical statistical procedures; therefore, it is a pairwise test that aims to detect significant differences between the behavior of two algorithms.

The null hypothesis for Wilcoxon's test is $H_0 : \theta_D = 0$; in the underlying populations represented by the two samples of results, the median of the difference scores equals zero. The alternative hypothesis is $H_1 : \theta_D \neq 0$, but also can be used $H_1 : \theta_D > 0$ or $H_1 : \theta_D < 0$ as directional hypothesis.

In the following, we describe the tests computations. Let d_i be the difference between the performance scores of the two algorithms on i-th out of N functions. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. Let R^+ be the sum of ranks for the functions on which the second algorithm outperformed the first, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i)$$

$$R^- = \sum_{d_i<0} \text{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \text{rank}(d_i)$$

Let T be the smallest of the sums, $T = \min(R^+, R^-)$. If T is less than or equal to the value of the distribution of Wilcoxon for N degrees of freedom (Table B.12 in Zar 1999), the null hypothesis of equality of means is rejected.

The obtaining of the p-value associated to a comparison is performed by means of the normal approximation for the Wilcoxon T statistic (Section VI, Test 18 in Sheskin 2003). Furthermore, the computation of the p-value for this test is usually included in well-known statistical software packages (SPSS, SAS, R, etc.).

A.3 The Friedman two-way analysis of variance by ranks

Friedman's test is used for answering this question: In a set of k samples (where $k \geq 2$), do at least two of the samples represent populations with different median

values? It is a non-parametric procedure employed in a hypothesis testing situation involving a design with two or more samples. It is the analogous of the repeated-measures ANOVA in non-parametrical statistical procedures; therefore, it is a multiple comparison test that aims to detect significant differences between the behavior of two or more algorithms.

The null hypothesis for Friedman's test is $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$; the median of the population i represents the median of the population j, $i \neq j, 1 \leq i \leq k$, $1 \leq j \leq k$. The alternative hypothesis is $H_1 : Not\ H_0$, so it is non-directional.

In the following, we describe the tests computations. It computes the ranking of the observed results for algorithm (r_j for the algorithm j with k algorithms) for each function, assigning to the best of them the ranking 1, and to the worst the ranking k. Under the null hypothesis, formed from supposing that the results of the algorithms are equivalent and, therefore, their rankings are also similar, the Friedman's statistic

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

is distributed according to χ_F^2 with $k-1$ degrees of freedom, being $R_j = \frac{1}{N}\sum_i r_i^j$, and N the number of functions. The critical values for the Friedman's statistic coincide with the established in the χ^2 distribution when $N > 10$ and $k > 5$. In a contrary case, the exact values can be seen in Sheskin (2003); Zar (1999).

Iman and Davenport (1980) proposed a derivation from the Friedman's statistic given that this last metric produces a conservative undesirably effect. The proposed statistic is

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2}$$

and it is distributed according to a F distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom.

Computation of the p-values given a χ^2 or F_F statistic can be done by using the algorithms in Abramowitz (1974). Also, most of the statistical software packages include it.

The rejection of the null hypothesis in both tests described above does not involve the detection of the existing differences among the algorithms compared. They only inform us about the presence of differences among all samples of results compared. In order to conducting pairwise comparisons within the framework of multiple comparisons, we can proceed with a post-hoc procedure. In this case, a control algorithm (maybe a proposal to be compared) is usually chosen. Then, the post-hoc procedures proceed to compare the control algorithm with the remain $k-1$ algorithms. In the following, we describe three post-hoc procedures:

- Bonferroni-Dunn's procedure (Zar 1999): it is similar to Dunnet's test for ANOVA designs. The performance of two algorithms is significantly different if the corre-

sponding average of rankings is at least as great as its critical difference (CD).

$$\mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

The value of q_α is the critical value of Q' for a multiple non-parametric comparison with a control (Table B.16 in Zar 1999).

- Holm (1979) procedure: for contrasting the procedure of Bonferroni-Dunn, we dispose of a procedure that sequentially checks the hypotheses ordered according to their significance. We will denote the p-values ordered by $p_1, p_2, \ldots,$ in the way that $p_1 \leq p_2 \leq \cdots \leq p_{k-1}$. Holm's method compares each p_i with $\alpha/(k-i)$ starting from the most significant p-value. If p_1 is below than $\alpha/(k-1)$, the corresponding hypothesis is rejected and it leaves us to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, we continue with the process. As soon as a certain hypothesis can not be rejected, all the remaining hypotheses are maintained as supported. The statistic for comparing the i algorithm with the j algorithm is:

$$z = (R_i - R_j) \Big/ \sqrt{\frac{k(k+1)}{6N}}$$

The value of z is used for finding the corresponding probability from the table of the normal distribution (p-value), which is compared with the corresponding value of α.

Holm's method is more powerful than Bonferroni-Dunn's and it does no additional assumptions about the hypotheses checked.

- Hochberg (1988) procedure: It is a step-up procedure that works in the opposite direction to Holm's method, comparing the largest p-value with α, the next largest with $\alpha/2$ and so forth until it encounters a hypothesis it can reject. All hypotheses with smaller p values are then rejected as well. Hochberg's method is more powerful than Holm's (Shaffer 1995).

When a p-value is within a multiple comparison it reflects the probability error of a certain comparison, but it does not take into account the remaining comparisons belonging to the family. One way to solve this problem is to report Adjusted P-Values (APVs) which take into account that multiple tests are conducted. An APV can be directly taken as the p-value of a hypothesis belonging to a comparison of multiple algorithms.

In the following, we will explain how to compute the APVs for the three post-hoc procedures described above, following the indications given in Wright (1992).

- Bonferroni APV_i: $\min\{v; 1\}$, where $v = (k-1)p_i$.
- Holm APV_i: $\min\{v; 1\}$, where $v = \max\{(k-j)p_j : 1 \leq j \leq i\}$.
- Hochberg APV_i: $\max\{(k-j)p_j : (k-1) \geq j \geq i\}$.

Appendix B: published average results of the CEC'2005 Special Session

Table 13 Average error rate obtained in CEC'2005 Special Session in dimension 10

Algorithm	f1	f2	f3	f4	f5	f6	f7	f8	f9
BLX-GL50	10^{-9}	10^{-9}	$5.705 \cdot 10^{2}$	10^{-9}	10^{-9}	10^{-9}	$1.172 \cdot 10^{-2}$	$2.035 \cdot 10$	1.154
BLX-MA	10^{-9}	10^{-9}	$4.771 \cdot 10^{4}$	$1.997 \cdot 10^{-8}$	$2.124 \cdot 10^{-2}$	1.49	$1.971 \cdot 10^{-1}$	$2.019 \cdot 10$	$4.379 \cdot 10^{-1}$
CoEVO	10^{-9}	10^{-9}	10^{-9}	10^{-9}	2.133	$1.246 \cdot 10$	$3.705 \cdot 10^{-2}$	$2.027 \cdot 10$	$1.919 \cdot 10$
DE	10^{-9}	10^{-9}	$1.94 \cdot 10^{-6}$	10^{-9}	10^{-9}	$1.59 \cdot 10^{-1}$	$1.46 \cdot 10^{-1}$	$2.04 \cdot 10$	$9.55 \cdot 10^{-1}$
DMS-L-PSO	10^{-9}	10^{-9}	10^{-9}	$1.885 \cdot 10^{-3}$	$1.138 \cdot 10^{-6}$	$6.892 \cdot 10^{-8}$	$4.519 \cdot 10^{-2}$	$2 \cdot 10$	10^{-9}
EDA	10^{-9}	10^{-9}	$2.121 \cdot 10$	10^{-9}	10^{-9}	$4.182 \cdot 10^{-2}$	$4.205 \cdot 10^{-1}$	$2.034 \cdot 10$	5.418
G-CMA-ES	10^{-9}	10^{-9}	10^{-9}	10^{-9}	10^{-9}	10^{-9}	10^{-9}	$2 \cdot 10$	$2.39 \cdot 10^{-1}$
K-PCX	10^{-9}	10^{-9}	$4.15 \cdot 10^{-1}$	$7.94 \cdot 10^{-7}$	$4.85 \cdot 10$	$4.78 \cdot 10^{-1}$	$2.31 \cdot 10^{-1}$	$2 \cdot 10$	$1.19 \cdot 10^{-1}$
L-CMA-ES	10^{-9}	10^{-9}	10^{-9}	$1.76 \cdot 10^{6}$	10^{-9}	10^{-9}	10^{-9}	$2 \cdot 10$	$4.49 \cdot 10$
L-SaDE	10^{-9}	10^{-9}	$1.672 \cdot 10^{-2}$	$1.418 \cdot 10^{-5}$	0.012	$1.199 \cdot 10^{-8}$	0.02	$2 \cdot 10$	10^{-9}
SPC-PNX	10^{-9}	10^{-9}	$1.081 \cdot 10^{5}$	10^{-9}	10^{-9}	$1.891 \cdot 10$	$8.261 \cdot 10^{-2}$	$2.099 \cdot 10$	4.02
Algorithm	f10	f11	f12	f13	f14	f15	f16	f17	f18
BLX-GL50	4.975	2.334	$4.069 \cdot 10^{2}$	$7.498 \cdot 10^{-1}$	2.172	$4 \cdot 10^{2}$	$9.349 \cdot 10$	$1.09 \cdot 10^{2}$	$4.2 \cdot 10^{2}$
BLX-MA	5.643	4.557	$7.43 \cdot 10$	$7.736 \cdot 10^{-1}$	2.03	$2.696 \cdot 10^{2}$	$1.016 \cdot 10^{2}$	$1.27 \cdot 10^{2}$	$8.033 \cdot 10^{2}$
CoEVO	$2.677 \cdot 10$	9.029	$6.046 \cdot 10^{2}$	1.137	3.706	$2.938 \cdot 10^{2}$	$1.772 \cdot 10^{2}$	$2.118 \cdot 10^{2}$	$9.014 \cdot 10^{2}$
DE	$1.25 \cdot 10$	$8.47 \cdot 10^{-1}$	$3.17 \cdot 10$	$9.77 \cdot 10^{-1}$	3.45	$2.59 \cdot 10^{2}$	$1.13 \cdot 10^{2}$	$1.15 \cdot 10^{2}$	$4 \cdot 10^{2}$
DMS-L-PSO	3.622	4.623	2.4001	$3.689 \cdot 10^{-1}$	2.36	4.854	$9.476 \cdot 10$	$1.101 \cdot 10^{2}$	$7.607 \cdot 10^{2}$
EDA	5.289	3.944	$4.423 \cdot 10^{2}$	1.841	2.63	$3.65 \cdot 10^{2}$	$1.439 \cdot 10^{2}$	$1.568 \cdot 10^{2}$	$4.832 \cdot 10^{2}$
G-CMA-ES	$7.96 \cdot 10^{-2}$	$9.34 \cdot 10^{-1}$	$2.93 \cdot 10$	$6.96 \cdot 10^{-1}$	3.01	$2.28 \cdot 10^{2}$	$9.13 \cdot 10$	$1.23 \cdot 10^{2}$	$3.32 \cdot 10^{2}$
K-PCX	$2.39 \cdot 10^{-1}$	6.65	$1.49 \cdot 10^{2}$	$6.53 \cdot 10^{-1}$	2.35	$5.1 \cdot 10^{2}$	$9.59 \cdot 10$	$9.73 \cdot 10$	$7.52 \cdot 10^{2}$
L-CMA-ES	$4.08 \cdot 10$	3.65	$2.09 \cdot 10^{2}$	$4.94 \cdot 10^{-1}$	4.01	$2.11 \cdot 10^{2}$	$1.05 \cdot 10^{2}$	$5.49 \cdot 10^{2}$	$4.97 \cdot 10^{2}$
L-SaDE	4.969	4.891	$4.501 \cdot 10^{-7}$	0.22	2.915	32	$1.012 \cdot 10^{2}$	$1.141 \cdot 10^{2}$	$7.194 \cdot 10^{2}$
SPC-PNX	7.304	1.91	$2.595 \cdot 10^{2}$	$8.379 \cdot 10^{-1}$	3.046	$2.538 \cdot 10^{2}$	$1.096 \cdot 10^{2}$	$1.19 \cdot 10^{2}$	$4.396 \cdot 10^{2}$
Algorithm	f19	f20	f21	f22	f23	f24	f25		
BLX-GL50	$4.49 \cdot 10^{2}$	$4.46 \cdot 10^{2}$	$6.893 \cdot 10^{2}$	$7.586 \cdot 10^{2}$	$6.389 \cdot 10^{2}$	$2 \cdot 10^{2}$	$4.036 \cdot 10^{2}$		
BLX-MA	$7.628 \cdot 10^{2}$	$8 \cdot 10^{2}$	$7.218 \cdot 10^{2}$	$6.709 \cdot 10^{2}$	$9.267 \cdot 10^{2}$	$2.24 \cdot 10^{2}$	$3.957 \cdot 10^{2}$		
CoEVO	$8.445 \cdot 10^{2}$	$8.629 \cdot 10^{2}$	$6.349 \cdot 10^{2}$	$7.789 \cdot 10^{2}$	$8.346 \cdot 10^{2}$	$3.138 \cdot 10^{2}$	$2.573 \cdot 10^{2}$		
DE	$4.2 \cdot 10^{2}$	$4.6 \cdot 10^{2}$	$4.92 \cdot 10^{2}$	$7.18 \cdot 10^{2}$	$5.72 \cdot 10^{2}$	$2 \cdot 10^{2}$	$9.23 \cdot 10^{2}$		
DMS-L-PSO	$7.143 \cdot 10^{2}$	$8.22 \cdot 10^{2}$	$5.36 \cdot 10^{2}$	$6.924 \cdot 10^{2}$	$7.303 \cdot 10^{2}$	$2.24 \cdot 10^{2}$	$3.657 \cdot 10^{2}$		
EDA	$5.644 \cdot 10^{2}$	$6.519 \cdot 10^{2}$	$4.84 \cdot 10^{2}$	$7.709 \cdot 10^{2}$	$6.405 \cdot 10^{2}$	$2 \cdot 10^{2}$	$3.73 \cdot 10^{2}$		
G-CMA-ES	$3.26 \cdot 10^{2}$	$3 \cdot 10^{2}$	$5 \cdot 10^{2}$	$7.29 \cdot 10^{2}$	$5.59 \cdot 10^{2}$	$2 \cdot 10^{2}$	$3.74 \cdot 10^{2}$		
K-PCX	$7.51 \cdot 10^{2}$	$8.13 \cdot 10^{2}$	$1.05 \cdot 10^{3}$	$6.59 \cdot 10^{2}$	$1.06 \cdot 10^{3}$	$4.06 \cdot 10^{2}$	$4.06 \cdot 10^{2}$		
L-CMA-ES	$5.16 \cdot 10^{2}$	$4.42 \cdot 10^{2}$	$4.04 \cdot 10^{2}$	$7.4 \cdot 10^{2}$	$7.91 \cdot 10^{2}$	$8.65 \cdot 10^{2}$	$4.42 \cdot 10^{2}$		
L-SaDE	$7.049 \cdot 10^{2}$	$7.13 \cdot 10^{2}$	$4.64 \cdot 10^{2}$	$7.349 \cdot 10^{2}$	$6.641 \cdot 10^{2}$	$2 \cdot 10^{2}$	$3.759 \cdot 10^{2}$		
SPC-PNX	$3.8 \cdot 10^{2}$	$4.4 \cdot 10^{2}$	$6.801 \cdot 10^{2}$	$7.493 \cdot 10^{2}$	$5.759 \cdot 10^{2}$	$2 \cdot 10^{2}$	$4.06 \cdot 10^{2}$		

References

Abramowitz, M.: Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables. Dover, New York (1974)

Auger, A., Hansen, N.: A restart CMA evolution strategy with increasing population size. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 1769–1776 (2005a)

Auger, A., Hansen, N.: Performance evaluation of an advanced local search evolutionary algorithm. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 1777–1784 (2005b)

Ballester, P.J., Stephenson, J., Carter, J.N., Gallagher, K.: Real-parameter optimization performance study on the CEC-2005 benchmark with SPC-PNX. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 498–505 (2005)

Bartz-Beielstein, T.: Experimental Research in Evolutionary Computation: The New Experimentalism. Springer, New York (2006)

Czarn, A., MacNish, C., Vijayan, K., Turlach, R., Gupta, R.: Statistical exploratory analysis of genetic algorithms. IEEE Trans. Evol. Comput. **8**(4), 405–421 (2004)

Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)

Gallagher, M., Yuan, B.: A general-purpose tunable landscape generator. IEEE Trans. Evol. Comput. **10**(5), 590–603 (2006)

García, S., Molina, D., Lozano, M., Herrera, F.: An experimental study on the use of non-parametric tests for analyzing the behaviour of evolutionary algorithms in optimization problems. In: Proceedings of the Spanish Congress on Metaheuristics, Evolutionary and Bioinspired Algorithms (MAEB'2007), pp. 275–285 (2007) (in Spanish)

García-Martínez, C., Lozano, M.: Hybrid real-coded genetic algorithms with female and male differentiation. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 896–903 (2005)

Hansen, N.: (2005). Compilation of results on the CEC benchmark function set. Tech. Report, Institute of Computational Science, ETH Zurich, Switzerland. Available as http://www.ntu.edu.sg/home/epnsugan/index_files/CEC-05/compareresults.pdf

Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika **75**, 800–803 (1988)

Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian J. Statist. **6**, 65–70 (1979)

Hooker, J.: Testing heuristics: we have it all wrong. J. Heuristics **1**(1), 33–42 (1997)

Iman, R.L., Davenport, J.M.: Approximations of the critical region of the Friedman statistic. Commun. Stat. **18**, 571–595 (1980)

Kramer, O.: An experimental analysis of evolution strategies and particle swarm optimisers using design of experiments. In: Proceedings of the Genetic and Evolutionary Computation Conference 2007 (GECCO'2007), pp. 674–681 (2007)

Liang, J.J., Suganthan, P.N.: Dynamic multi-swarm particle swarm optimizer with local search. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 522–528 (2005)

Molina, D., Herrera, F., Lozano, M.: Adaptive local search parameters for real-coded memetic algorithms. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 888–895 (2005)

Moreno-Pérez, J.A., Campos-Rodríguez, C., Laguna, M.: On the comparison of metaheuristics through non-parametric statistical techniques. In: Proceedings of the Spanish Congress on Metaheuristics, Evolutionary and Bioinspired Algorithms (MAEB'2007), pp. 286–293 (2007) (in Spanish)

Morse, D.T.: Minsize2: a computer program for determining effect size and minimum sample size for statistical significance for univariate, multivariate, and nonparametric tests. Educ. Psychol. Meas. **59**(3), 518–531 (1999)

Noether, G.E.: Sample size determination for some common nonparametric tests. J. Am. Stat. Assoc. **82**(398), 645–647 (1987)

Ortiz-Boyer, D., Hervás-Martínez, C., García-Pedrajas, N.: Improving crossover operators for real-coded genetic algorithms using virtual parents. J. Heuristics **13**, 265–314 (2007)

Ozcelik, B., Erzurumlu, T.: Comparison of the warpage optimization in the plastic injection molding using ANOVA, neural network model and genetic algorithm. J. Mater. Process. Technol. **171**(3), 437–445 (2006)

Patel, J.K., Read, C.B.: Handbook of the Normal Distribution. Dekker, New York (1982)

Pošík, P.: Real-parameter optimization using the mutation step co-evolution. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 872–879 (2005)

Qin, A.K., Suganthan, P.N.: Self-adaptive differential evolution algorithm for numerical optimization. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 1785–1791 (2005)

Rojas, I., González, J., Pomares, H., Merelo, J.J., Castillo, P.A., Romero, G.: Statistical analysis of the main parameters involved in the design of a genetic algorithm. IEEE Trans. Syst. Man Cybern. Part C **32**(1), 31–37 (2002)

Rônkkônen, J., Kukkonen, S., Price, K.V.: Real-parameter optimization using the mutation step co-evolution. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 506–513 (2005)

Shaffer, J.P.: Multiple hypothesis testing. Annu. Rev. Psychol. **46**, 561–584 (1995)

Sinha, A., Tiwari, S., Deb, K.: A population-based, steady-state procedure for real-parameter optimization. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 514–521 (2005)

Suganthan, P.N., Hansen, N., Liang, J.J., Deb, K., Chen, Y.P., Auger, A., Tiwari, S.: (2005). Problem definitions and evaluation criteria for the CEC 2005 Special Session on Real Parameter Optimization. Tech. Report, Nanyang Technological University. Available as http://www.ntu.edu.sg/home/epnsugan/index_files/CEC-05/Tech-Report-May-30-05.pdf

Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. CRC Press, Boca Raton (2003)

Whitley, D.L., Beveridge, R., Graves, C., Mathias, K.E.: Test driving three 1995 genetic algorithms: new test functions and geometric matching. J. Heuristics **1**(1), 77–104 (1995)

Whitley, D.L., Rana, S., Dzubera, J., Mathias, K.E.: Evaluating evolutionary algorithms. Artif. Intell. **85**(1–2), 245–276 (1996)

Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997)

Wright, S.P.: Adjusted p-values for simultaneous inference. Biometrics **48**, 1005–1013 (1992)

Yuan, B., Gallagher, M.: On building a principled framework for evaluating and testing evolutionary algorithms: a continuous landscape generator. In: Proceedings of the 2003 Congress on Evolutionary Computation (CEC2003), pp. 451–458 (2003)

Yuan, B., Gallagher, M.: Experimental results for the Special Session on Real-Parameter Optimization at CEC 2005: a simple, continuous EDA. In: Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC'2005), pp. 1792–1799 (2005)

Zar, J.H.: Biostatistical Analysis. Prentice Hall, Englewood Cliffs (1999)