OPTIMIZATION TECHNIQUES

STATISTICAL ANALYSIS OF EVOLUTIONARY ALGORITHMS

Antonio LaTorre (atorre@fi.upm.es)

Motivations

- There are plenty of algorithms that can be used to solve different problems
- Imagine that we have the results for all those algorithms on that problem
- How can we assess which is the outstanding algorithm for that problem?
- And if we are comparing the algorithms on multiple problems?

Relevant work

- Three types of work in the analysis of experiments within the field of EC:
  - Design of test problems
  - **Statistical analysis of the results**
    - Average and standard deviation of multiple executions is definitely **NOT enough!!**
  - Experimental design
    - Parameter tuning, etc.

Preliminaries

- Some useful Definitions
  - **Statistical test:** procedure to check if one hypothesis holds by analyzing some data distribution(s) (normality of one distribution, comparison of two distributions, etc.)
  - **p-value:** result reported by a statistical test which expresses the probability for a hypothesis to be true
  - **Confidence level (α):** threshold chosen to reject the hypothesis checked by the statistical test
    - values can range from 1% to 10%

Preliminaries

- In order to easily introduce the concepts of statistical validation of EAs, we will use a practical example with data from the Special Session on Continuous Optimization of CEC 2005
  - 11 participating algorithms
  - 25 test functions
    - 5 unimodal
    - 20 multimodal
  - 25 executions of each algorithm
    - We record error rate (difference with optimum)
  - Dimension D = 10 for all the functions
  - 100,000 Fitness Evaluations (FEs) allowed for each execution
  - Stop criterion: maximum number of FEs or 10^{-8} precision reached

Parametric Tests

- Parametric Tests make some assumptions about the data under consideration
  - Real-valued data
    - Fulfilled, as we record fitness values
  - Independence of events which generated data
    - Obvious in the case of EAs, as executions are run independently with different random seeds
  - **Normality of the distribution of data**
  - **Heterocedasticity of variances**

Parametric Tests

- ***Normality***
  - Results follow a Gaussian distribution with a certain average and variance
  - Three normality tests
    - **Kolmogorov-Smirnov:** compares accumulated distribution of observed data and Gaussian distribution
    - **Shapiro-Wilk:** Analyzes the observed data to compute the level of symmetry and kurtosis to compare it to a Gaussian distribution
    - **D'Agostino-Pearson:** Computes the skewness and kurtosis of the distribution to see how far it is from the Gaussian distribution

Parametric Tests

- ***Heteroscedasticity***
  - Checks if k samples present homogeneity of variances (homoscedasticity)
  - Two tests
    - Levene's Test (preferable when the distribution is not normal)
    - Bartlett's Test

Single problem analysis

- We are going to check normality and heteroscedasticity for two algorithms (BLX-GL50 and BLX-MA) in the 25 functions (25 executions per function)
  - Three normality tests
  - Only Leven's test for heteroscedasticity
- Low levels of p-value indicate a non-normal distribution
  - Significance level $\alpha = 0.05$

Single problem analysis

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)		

Single problem analysis

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)

Single problem analysis

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)		

Single problem analysis

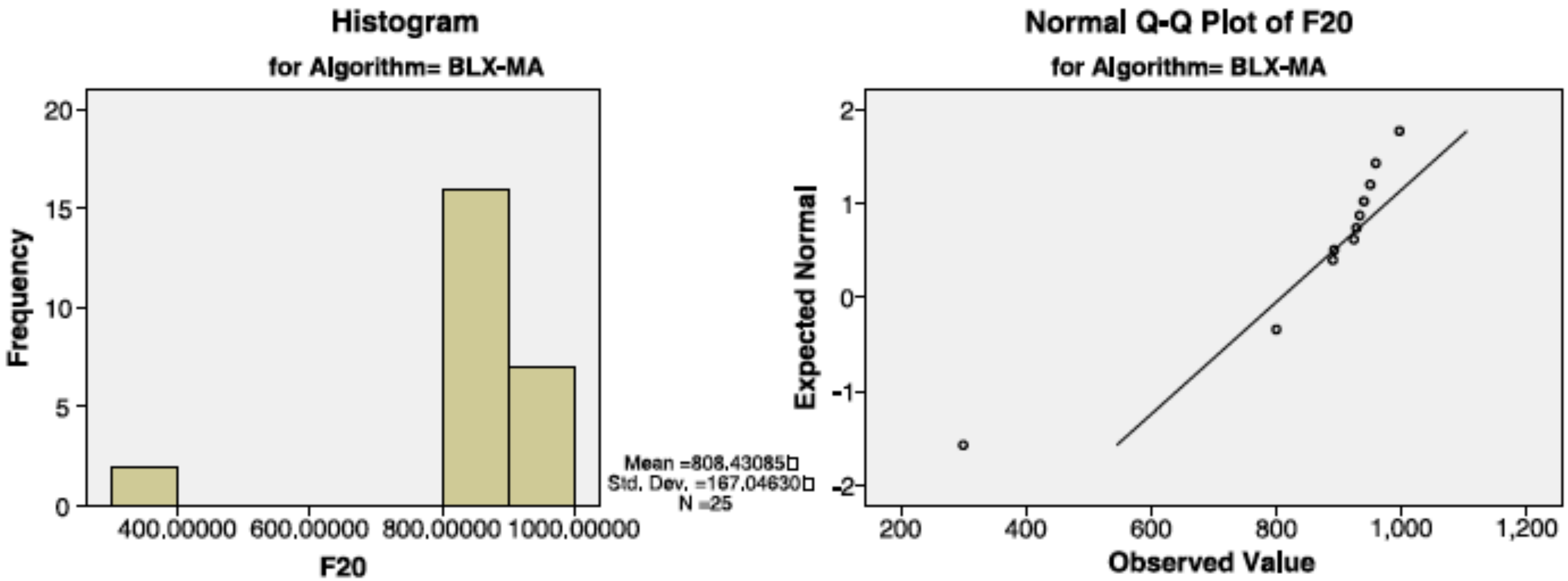

Fig. 1 Example of non-normal distribution: Function f20 and BLX-GL50 algorithm: Histogram and Q-Q Graphic

Single problem analysis

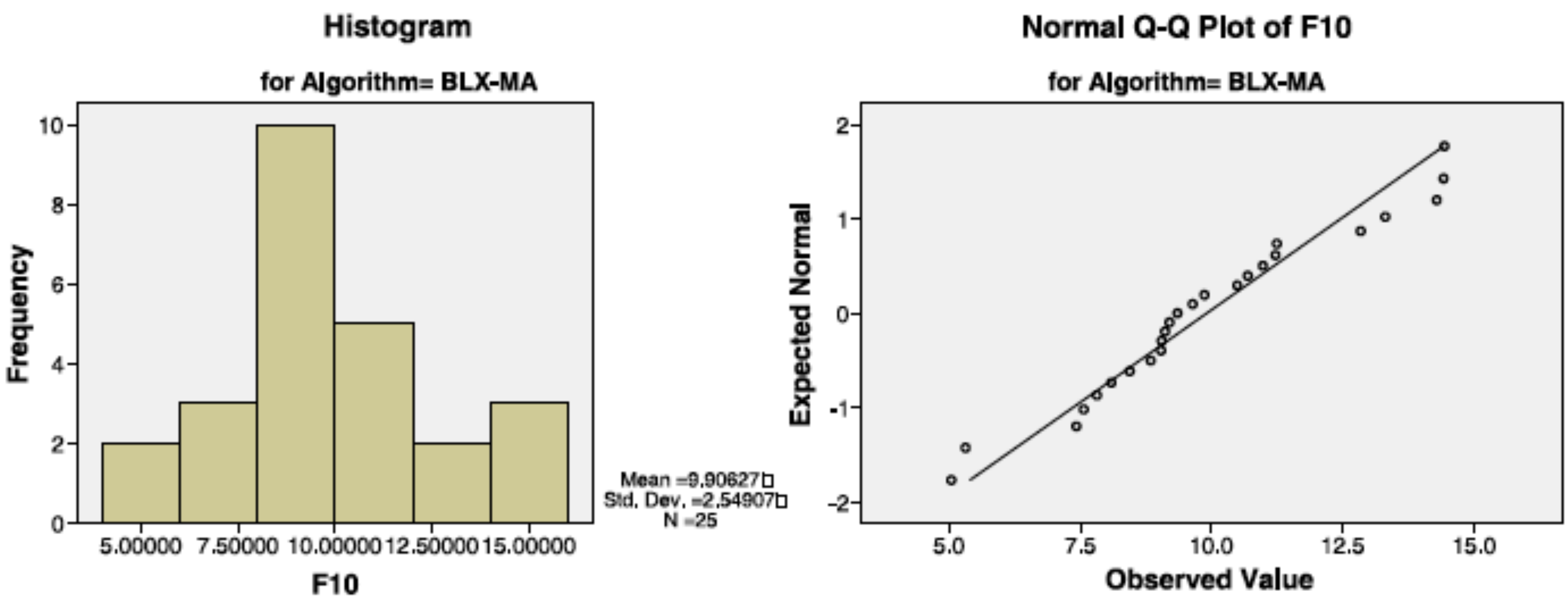

Fig. 2 Example of normal distribution: Function f10 and BLX-MA algorithm: Histogram and Q-Q Graphic

Single problem analysis

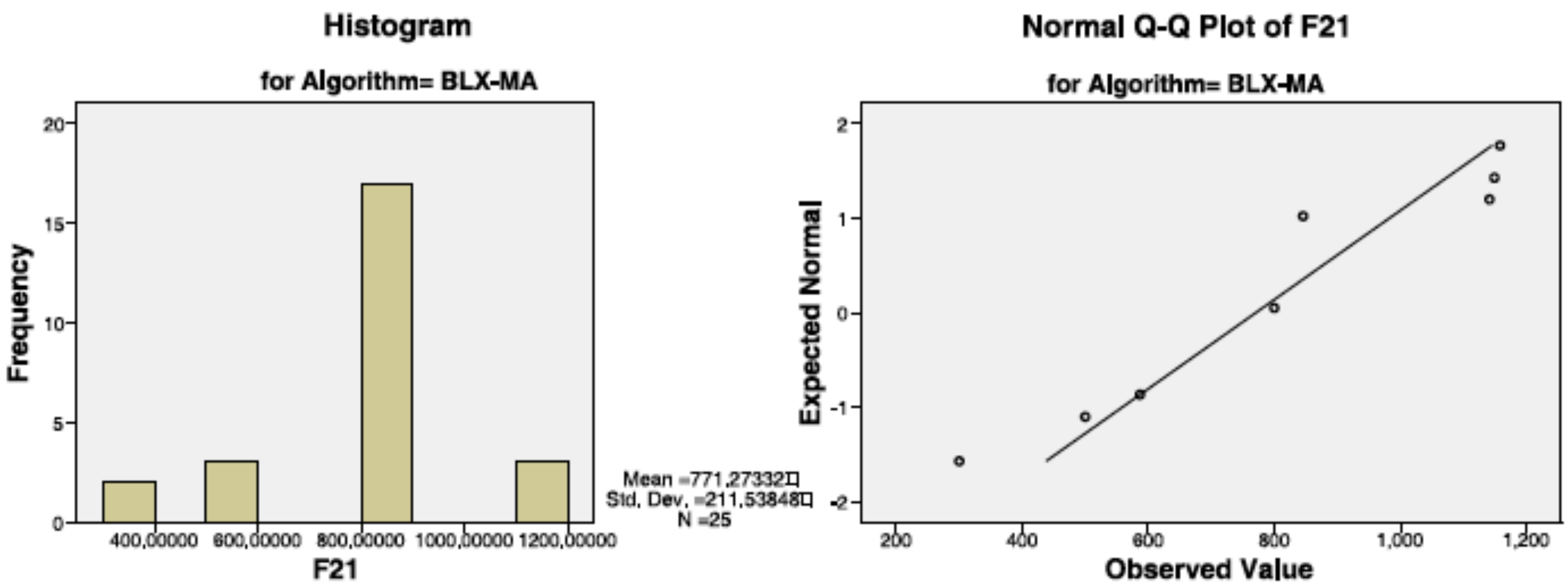

Fig. 3 Example of a special case: Function f21 and BLX-MA algorithm: Histogram and Q-Q Graphic

Single problem analysis

- There are some functions for which normality tests give contradictory results
  - It depends on the input data (size and distribution)
  - Normally, researchers choose the one which supports their hypothesis…
  - It should be carefully chosen and, in case of large discrepancies, results should be taken with care

Single problem analysis

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)		

Single problem analysis

- Normality and homostedasticity conditions are not fulfilled in many functions
- A researcher may think that this is not that important and use parametric instead on those functions
- We will see an example of what happens when this is done

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

Single problem analysis

- In three functions there are great differences
  - f4: Wilcoxon test considers that both algorithms behave differently, whereas t-test says the opposite
  - f15: opposite situation
  - f22: both p-values are greater than 0.05, but very different among them

Single problem analysis

- What can we do to avoid these problems?
  - **Use non-parametric tests if safety conditions are not fulfilled!!!**

- Yes, ok, but what else?
  - Conduct more executions of your problem in order to have more information
  - Transform your data to obtain normal distributions (logarithm, square root, etc.)
  - Skip outliers (use with great care)

Multiple problems analysis

- It is actually not much different to single problem analysis
- We need a mean to average the results of multiple problems
  - Normally, the average for each problem is considered
  - It is preferably that the same number of repetitions is done for each problem and algorithm
- We will consider the same two algorithms on the whole set of CEC 2005 functions

Multiple problems analysis

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)

Multiple problems analysis

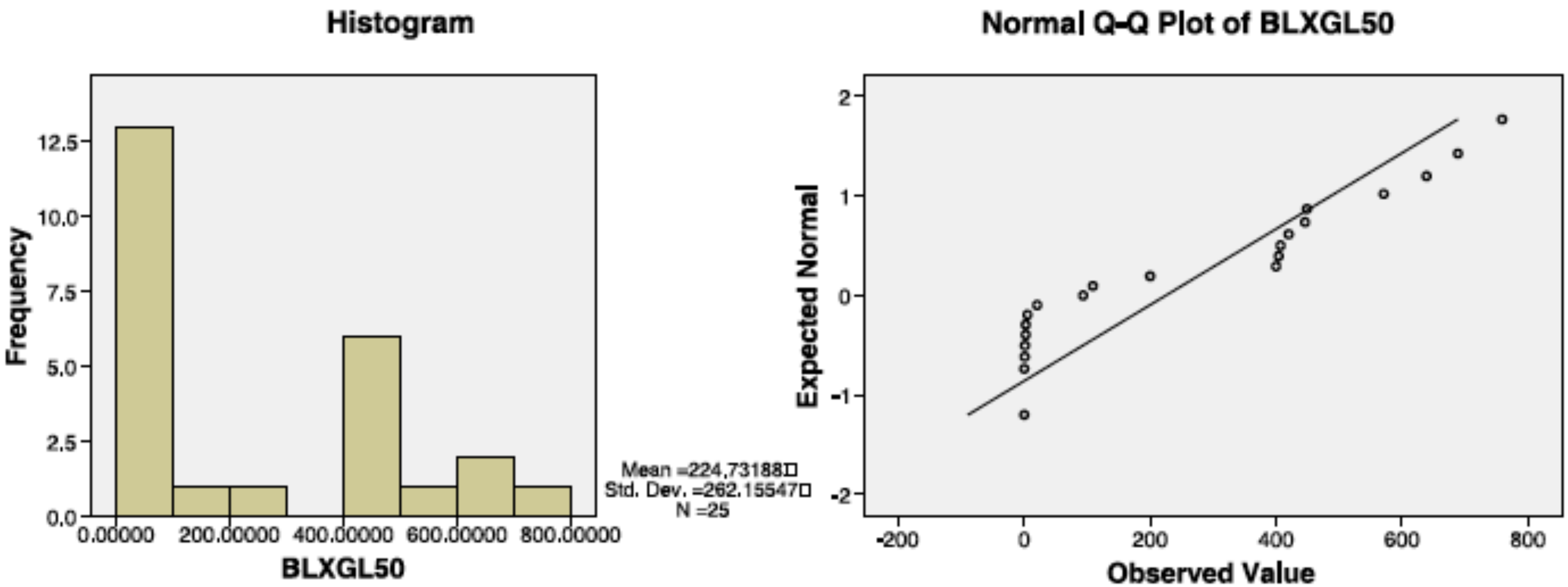

Fig. 4 BLX-GL50 algorithm: Histogram and Q-Q Graphic

Multiple problems analysis

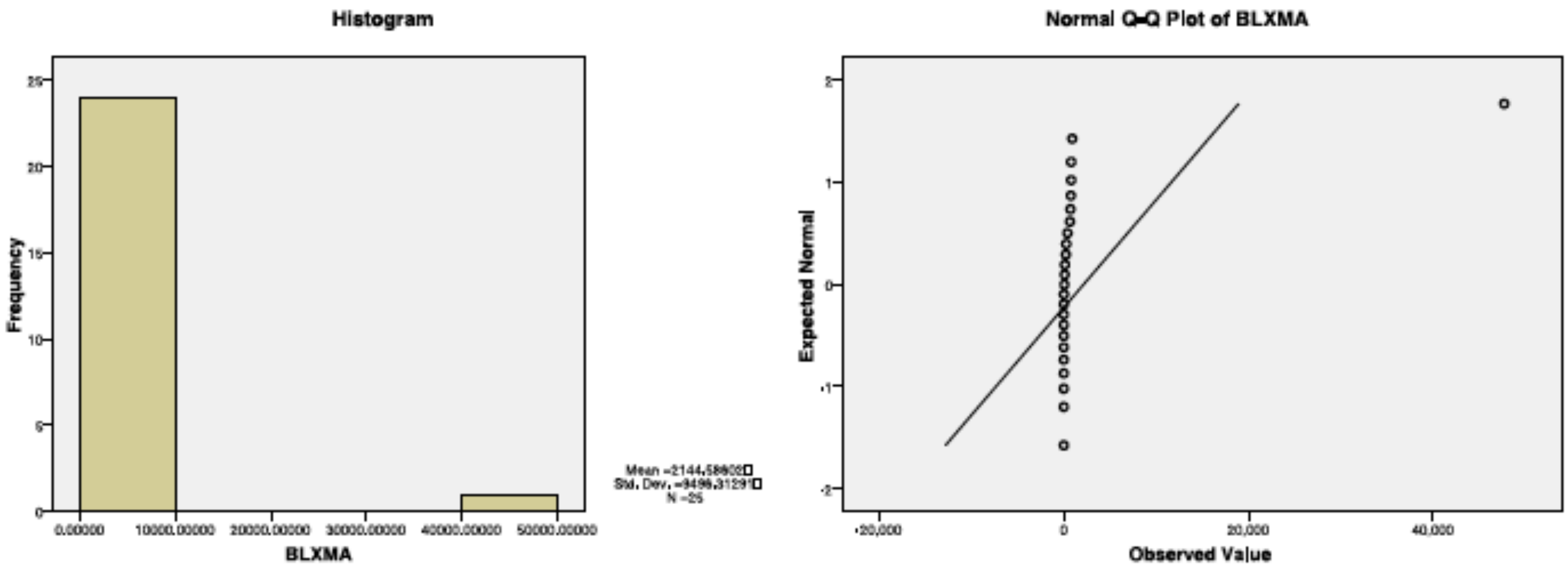

Fig. 5 BLX-MA algorithm: Histogram and Q-Q Graphic

Multiple problems analysis

- None of the conditions is fulfilled
  - We can not enlarge the number of results in multiple problems analysis (the number of results is the number of problems)
  - We can not discard “outliers” without biasing the result of the test
  - No transformation is likely to work with multiple problems

Use non-parametric tests!!!

Non-Parametric Tests

- For the introduction of the different tests involved in the statistical analysis, we will use the 11 algorithms and 25 functions of the CEC 2005 Special Session
- Functions will be grouped into two groups
  - "Difficult" functions: f15-f25
  - All functions: f1-f25

- The study will try to compare the algorithm with the lowest average error rate of the Special Session: G-CMA-ES

Friedman and Iman-Davenport Tests

- These tests are used to check if there are significant differences in the distribution of several sets of data (results of different algorithms)
  - Friedman Test: compares the median of the distributions
  - Iman-Davenport Test: Derivation from the Friedman test to correct the conservative behavior of the first one under some situations
- Given a set of *k* algorithms and *N* functions:

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

$$\text{with} \quad R_j = \frac{1}{N}\sum_i r_j^i$$

Friedman Statistic

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

Iman-Davenport Correction

Friedman and Iman-Davenport Tests

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

Friedman and Iman-Davenport Tests

- If the result of the Friedman / Iman-Davenport Test is **significant** for data coming from different distributions, we should then use other tests to test the hypothesis of a **reference algorithm** (normally the one with the best average ranking R_i) being better than the other ones
  - In our example, this is clearly true, so we can proceed to the next step

Bonferroni-Dunn's Test

- Checks if the performance of two algorithms is significantly different
- Normally, the algorithm with the best average ranking is compared with the other ones
- The difference is significant if the corresponding ranking is greater than the critical difference value, which is computed as follows

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

Bonferroni-Dunn's Test

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**

Bonferroni-Dunn's Test

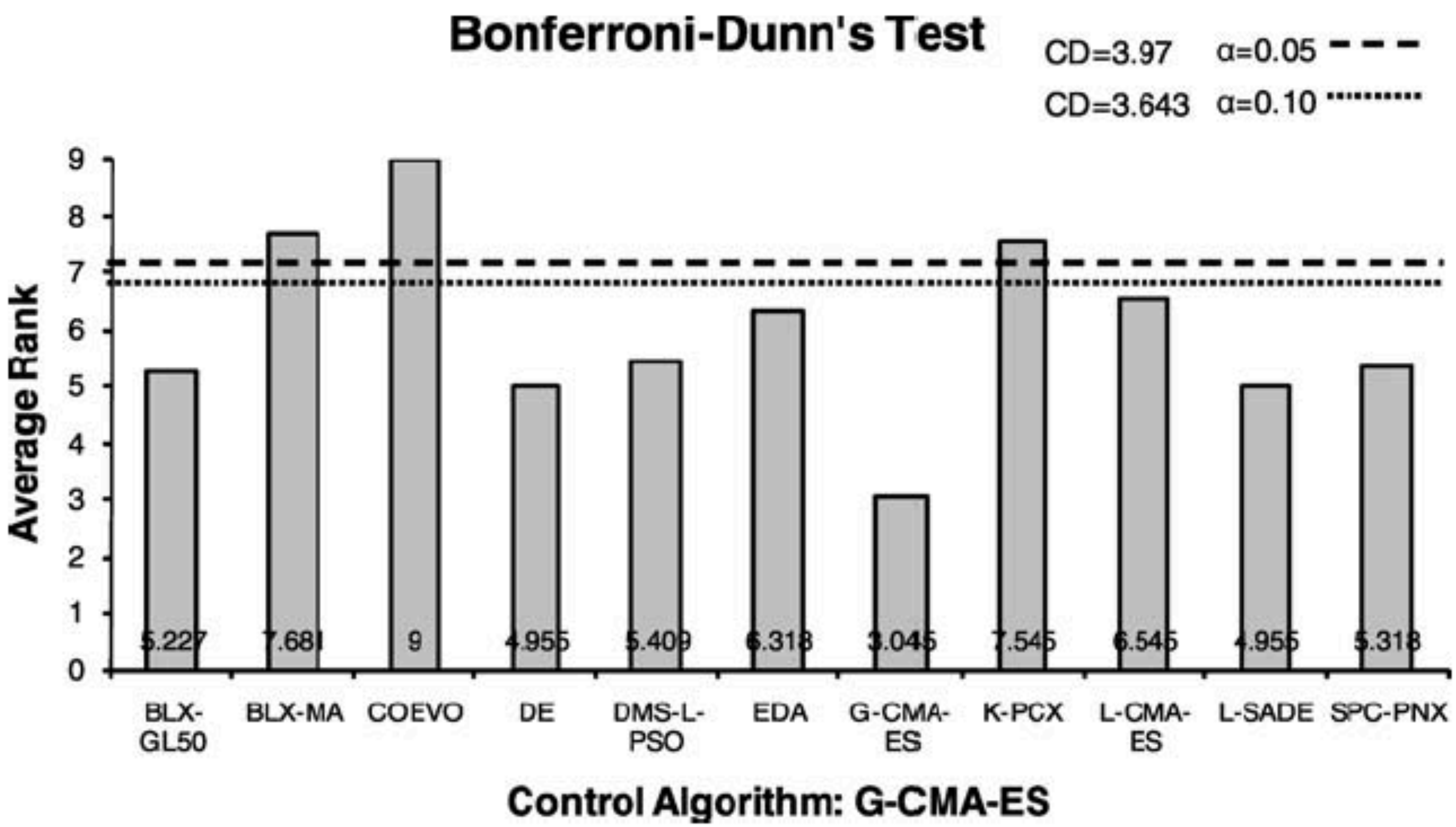

Fig. 6 Bonferroni-Dunn's graphic corresponding to the results for f15–f25

Bonferroni-Dunn's Test

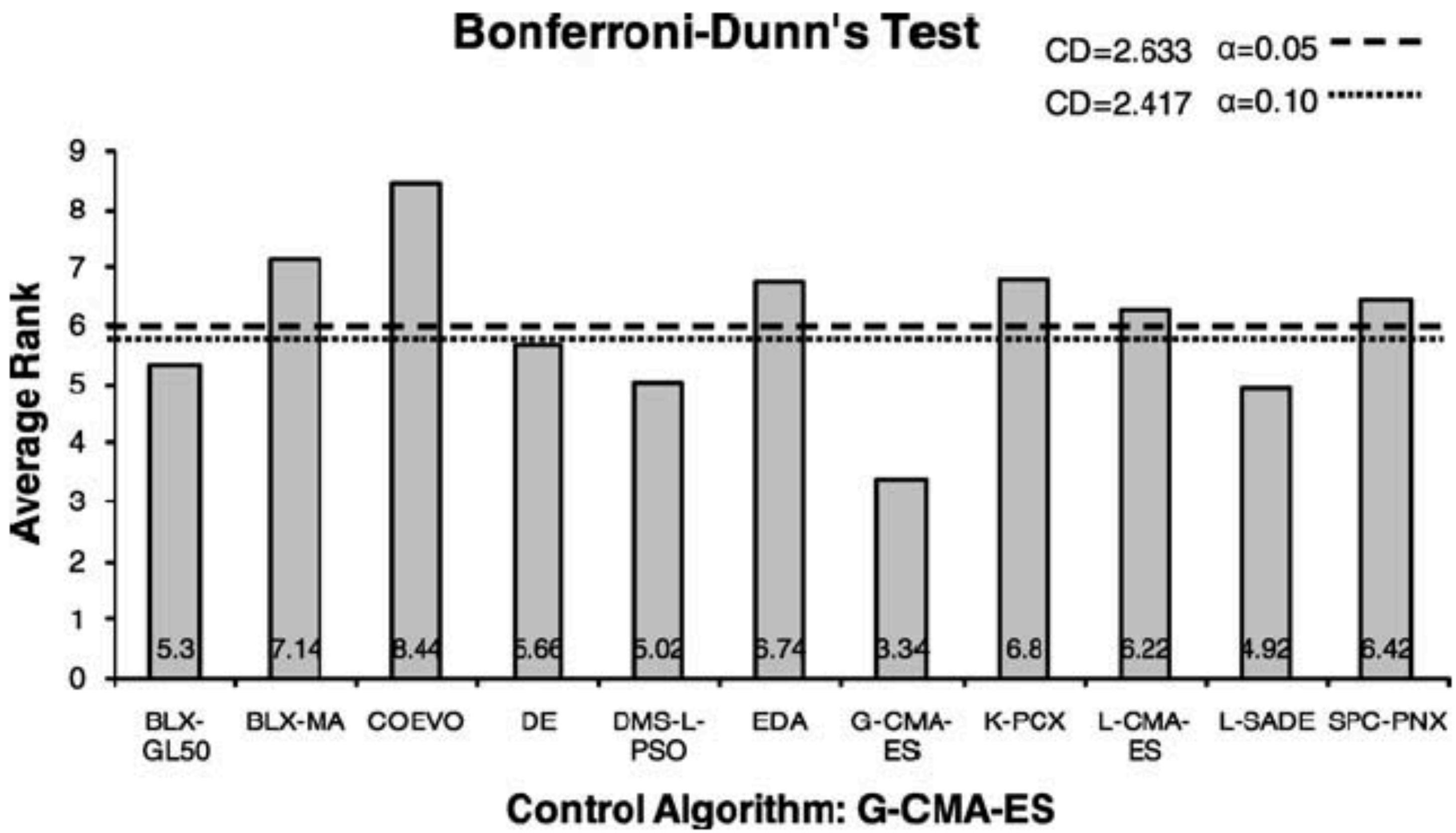

Fig. 7 Bonferroni-Dunn's graphic corresponding to the results for f1–f25

Holm Test

- Holm's procedure is more powerful than Bonferroni-Dunn's Test
- It is an iterative process that sequentially checks the hypotheses according to their significance
- p-values are ordered such as $p_1 \leq p_2 \leq \ldots \leq p_{k-1}$
- Each p_i is compared with $\alpha/(k - i)$, starting by p_1
- If p_1 is below $\alpha/(k - 1)$, then we continue with p_2, and so on
- As soon as one hypothesis can not be rejected, the remaining hypothesis remain supported
- The statistical used for comparing algorithms is:
- This value is used to obtain the p-value from the Normal distribution

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

Hochberg Test

- It works in the opposite sense of Holm's method
- It compares the largest p-value with α, the next largest with $\alpha/2$, $\alpha/3$… and so on until it encounters one hypothesis that it can reject
- All hypotheses with smaller p-values are then rejected as well
- Some studies state that Hochberg test is more powerful than Holm's

Adjustment of p-values

- A p-value reflects the probability error of a certain comparison, but it does not take into account the remaining comparisons
- An Adjusted P-Value (APV) needs to be computed
  - 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: $\min\,\{(k - j)\, p_j : (k - 1) \leq j \leq i\}$
- These adjusted p-values are used for the study (and reported on the tables)

Adjusted p-values

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

Adjusted p-values

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

Comparison of the different tests

- We consider G-CMA-ES as the control algorithm
- f15-f25
  - α =0.05 : both Holm's and Hochberg's test agree that G-CMA-ES is better than 3 algorithms
  - α =0.10 : both Holm's and Hochberg's test agree that G-CMA-ES is better than 4 algorithms (one more than Bonferroni's)
- f1-f25 :
  - α =0.05 : both Holm's and Hochberg's test agree that G-CMA-ES is better than 6 algorithms
  - α =0.10 : Holm's test gives significant results for 7 algorithms (one more than Bonferroni's). Hochberg's test gives significant results for all the 10 algorithms

Pairwise comparison (Wilcoxon Test)

- Considers only two algorithms at each comparison
- Aims to detect if there are significant differences between the behavior of two algorithms (equivalent to the t-test in parametrical tests)
- The null hypothesis is that the difference of the medians of both distributions is zero
  - Alternative hypothesis can be defined in both senses

Pairwise comparison (Wilcoxon Test)

- Being d_i the difference in performance for the two algorithms in function i we compute the Wilcoxon p-value in the following way

$$R^+ = \sum_{d_i > 0} rank(d_i)$$

$$R^- = \sum_{d_i < 0} rank(d_i)$$

- The Wilcoxon statistic is $T = \min(R^+, R^-)$, and the p-value is obtained from the appropriate table of approximations

Family Wise Error Rate (FWER)

- If we want to obtain relevant conclusions from pairwise comparison we must consider the Family Wise Error Rate (FWER)
  - Accumulated error coming from the combination of multiple pairwise comparisons
  - It is the probability of making one or more false discoveries when performing multiple pairwise algorithms

Wilcoxon Test

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

Wilcoxon Test

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

Family Wise Error Rate (FWER)

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

Wilcoxon Test

- For functions f15-f25, G-CMA-ES is better than BLX-GL50, BLX-MA, CoEVO, DE, EDA, K-PCX, L-CMA-ES and SPC-PNX with a p-value of:

$$p = 1 - ((1 - 0.001) \cdot (1 - 0.001) \cdot (1 - 0.000) \cdot (1 - 0.009) \cdot (1 - 0.001) \cdot (1 - 0.004) \cdot (1 - 0.003) \cdot (1 - 0.000)) = 0.018874$$

- And for functions f1-f25:

$$p = 1 - ((1 - 0.009) \cdot (1 - 0.016) \cdot (1 - 0.016) \cdot (1 - 0.028) \cdot (1 - 0.013) \cdot (1 - 0.016) \cdot (1 - 0.026) \cdot (1 - 0.007)) = 0.123906$$

Final considerations

- First step, detecting if there are differences in the means (Friedman or Iman-Davenport)
- If these algorithms detect differences, the **Holm** procedure should be used instead of the Bonferroni (it controls the FWER)
- **Hochberg's** procedure can be more precise than Holm's and may be used simultaneously
- Thumb rule to determine if non-parametric tests can be used safely: minimum number of samples **$N = a \cdot k$**, being k the **number of algorithms**