Biaxial Box Plots and Ordered Trial Ranks for Visualizing Large Sets of Experimental Results

Kent McClymont
University of Exeter
Streatham Campus, North Park Road
Exeter, UK
+44 (0)1392 723628
k.mcclymont@exeter.ac.uk

ABSTRACT
This paper presents a novel method for visualizing large experimental datasets called a Biaxial Box Plot which provides both an easily read general impression of the results that highlights performance trends whilst also allowing for careful comparison of individual results. The Biaxial Box Plot is compared against heatmaps and traditional box plots where it is argued that the new method provides a suitable combination of the two existing methods. In addition, a novel ranking method is presented called the Ordered Trial Rank (OTR) that is designed for use with results that contain a large number of related sets of samples – e.g. a group of algorithm performance results on the same problem. The OTR is compared against simple median and standard deviation scores and shown to provide a better statistical distinction between the sets of results. Both methods are presented in the context of EA experimental research but can be applied more generally to data with two orthogonal group that can be combined to create a matrix of numeric data sets.

Categories and Subject Descriptors
H.5.0 [**Information Interfaces and Presentation**]: General.

General Terms
Algorithms, Experimentation Theory.

Keywords
Visualization, Sorting, Algorithms, Experimental Design.

1. INTRODUCTION
Optimization techniques are becoming ever more general in their application. Algorithms in Evolutionary Computation, such as Evolutionary Algorithms (EAs), are widely applied across a number of problem domains and under a vast array of conditions [1]. In many cases, specific variations on well-known algorithms such as NSGA-II [2] are used, tailoring these general algorithm frameworks to more effectively solve specific problems. While, the trend for ever more abstracted search methodologies has recently led to the emergence of two fields of research: Hyper-heuristics [3] and Memetic Algorithms [4].

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.

The large number of EA variants and the increasing number of proposed methods for hyper-heuristics and Memetic Algorithms (for example) has led to an explosion in number of algorithms that need to be compared to one another when conducting optimization studies. It is no longer acceptable to simply compare against NSGA-II or other "benchmark" algorithms. Furthermore, the generality of these algorithms means that each method can be applied to increasing numbers of optimization problems. Consequently, methods for visualizing the large datasets generated by repeated runs of the algorithms (e.g., single objective, generational distance and hypervolume value results) are needed that both provide a good overview of each methods performance whilst also providing enough detail to allow for careful comparison between individual sets of results.

This paper presents a novel method for visualizing large experimental datasets in a matrix structure called a Biaxial Box Plot which provides both an easily read general impression of the results and highlights performance trends whilst also allowing for careful comparison of individual results. The method mirrors the results in both the horizontal and vertical to allow for an analysis of results across rows and down columns. The Biaxial Box Plot is compared against heatmaps and traditional box plots where it is argued that the new method provides a suitable combination of the two existing methods.

A novel ranking method is also presented called the Ordered Trial Rank (OTR) that is tailored specifically for use with sets of results where there are a large number of related sets of samples – e.g. a group of algorithm performance results on the same problem. The OTR is compared against simple median and standard deviation scores and shown to provide a better distinction between the sets of results. Both methods are presented in the context of EA experimental research but can be applied more generally to data with two orthogonal group that can be combined to create a matrix of numeric data sets.

The source code is available in Java and C#.NET and can be downloaded from http://people.exeter.ac.uk/km314/.

2. BIAXIAL BOX PLOT
Consider, for example, a study of optimization methods which compares the set of algorithms that are applicable to real-valued optimization problems by applying each algorithm to every benchmark problem in the DTLZ [5], WFG [6], and LZ09 [7] test problem suites. Such a study would be useful as it would indicate which sets of problems each algorithm is well suited to as well as give an idea of the relative quality of the algorithms across the set of problems. The number of available optimization approaches is

very large and even a small subset of these methods would still yield significant number of results.

2.1 Heatmaps

A simple approach to representing these results would be to use a heatmap. Heatmaps are matrix structures that use cells to represent data associated with specific row/cell combinations. For example, a 4x3 heatmap would contain 12 cells representing each combination of row and column elements. The cells are coloured using a colour scale based on the value in each cell. The colour range is commonly scaled between the worst and best value in the matrix. This example is illustrated in Figure 1.

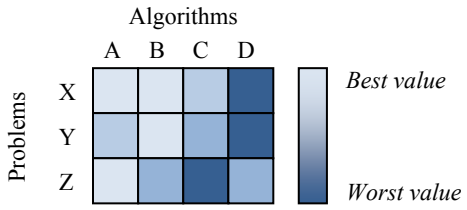

Figure 1. Illustration of a heatmap of multiple algorithm results on a set of problems

However, due to the stochastic nature of most optimizing algorithms, the results generated by a single optimization run are rarely representative of the algorithms' average performance. Running an algorithm over a number of trial runs is common practice in optimization studies and provides means of assessing both the average performance of an algorithm as well as the consistency of results.

Continuing with the example above, each cell in the heatmap would then represent a set of trial run results for an algorithm on a problem. A sensible means of representing this data would be to take the average result and use that to generate the colour value. However, this introduces data loss and the distribution of results is lost in this means of visualization.

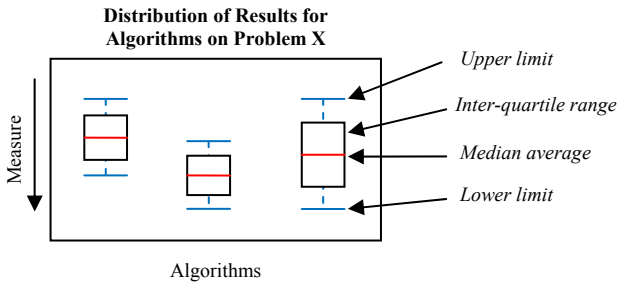

Figure 2. Illustration of a box plot of multiple algorithm results on a single problems

2.2 Box Plots

While heatmaps provide a reasonable means of depicting average results it is impossible to demonstrate the associated variance which is observed after a number of trial runs on of each algorithm on each problem. Box plots (illustrated in Figure 2) display both median and the variability in performance of set of trial run results and are well suited to demonstrating results from a set of algorithms on one problem (or vice versa) but do not easily demonstrate large sets of results.

However, the increase in information presented in the box plot figure format is accompanied by some negative consequences. Box plots are not an effective visualization method for showing the larger matrices of results depicted in heatmaps. When showing sets of boxplots together, they become difficult read and distinguish between. Heatmaps also allow for easy comparison of results in both axes – giving an overall picture of an algorithm's performance on a large number of problems as well as the set of algorithms on one problem.

2.3 Bagplots (Bivariate Boxplots)

In [9], Rousseeuw et al. define a form of visualization called a bagplot. Rousseeuw et al. identify the limitations of the univariate box plot method and provide a means for visualization bivariate samples. The method extends the conceptual depiction of median and range values and uses a shaded region (the 'bag') to highlight and encase the 50% central samples and an outer shaded region (the 'fence') to highlight and separate the inliers from the outliers.

In many respects, the bagplot is similar to the Biaxial Box Plot shown below, extending the function of the box plot for wider use. However, there are some crucial differences. A basic bagplot visualizes a single set of samples over two variables while the Biaxial Box Plot shows paired sets of samples from a single variable (i.e., each cell in the Biaxial Box Plot matrix represents a univariate distribution).

In addition, the bagplot is drawn over a two dimensional real space while the Biaxial Box Plot is an unordered matrix of sets. Rousseeuw et al. [9] do demonstrate how the bagplots can be arranged in a matrix format to display a single series or groups of results. The same series is used for both dimensions in the matrix unlike the biaxial box plot which pairs two series or groups.

The bagplot provides an excellent means for visualizing scalar bivariate data. The Biaxial Box Plot, in contrast, is a means for displaying paired series or groups of univariate data, providing a different extension of the boxplot format.

2.4 Biaxial Box Plot

Using the matrix format of heatmaps as a template, it is possible to restructure box plots to show less detailed information but across a wider set of data, as illustrated in Figure 3.

2.4.1 Average and Variance

Each cell in the matrix is used to represent the results from multiple trial runs of one algorithm on one problem in a box plot inspired depiction of results. The median average value is represented by a central line which is shown in both horizontal and vertical axes. The intersection of the two lines is surrounded by a dashed rectangle which shows the inter-quartile range – again shown for both axes. The mirroring/repetition of the results allows the box plots to be read along a row (for all algorithms on one problem) or down a column (for one algorithm on all problems).

The red median value lines are drawn to a distance of two times the bound (upper or lower) in either direction, with a minimum length of 1/3 of the cell width/height to ensure they are suitably visible. The lines are also limited to within the boundary of each cell. The lengths of the median lines do not give an indication of value but are extended to make the median value visible.

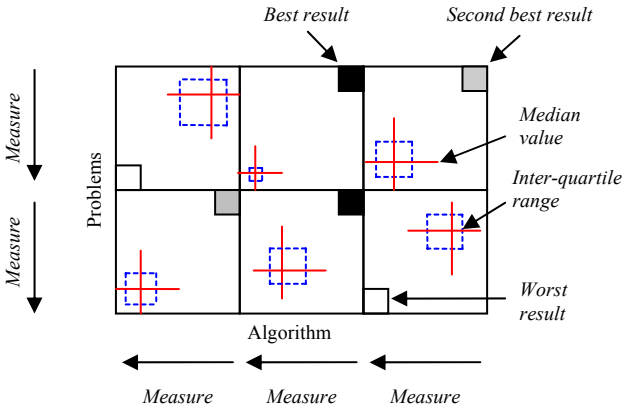

Figure 3. Illustration of the Biaxial Box Plot of multiple algorithm results on multiple of problems

Similarly to the heatmap, the results are normalized in the range [0, 1], unless the results are already normalized within this range. The Ordered Trial Rank, described later in Section 3, produces values in the range [0, 1] and so is ideally suited to this representation. In cases where the results have been normalized, the upper and lower values for each row should be appended to the right end of the row, illustrated in Figure 4.

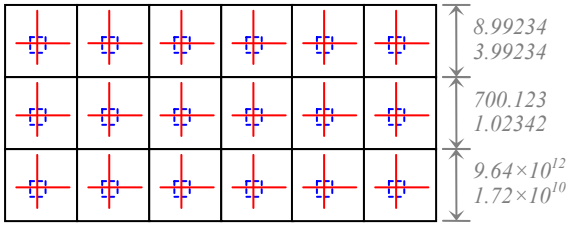

Figure 4. Biaxial Box Plot with ranges.

2.4.2 Markers

In addition to providing the median and variance results, the matrix is labeled with markers to indicate the best result for each row (again shown in Figure 3). This is shown by a black filled rectangle in the top right of the cell. The second best result is shown by a gray filled rectangle, again in the top right of the cell. The worst result is indicated by an unfilled (white) rectangle in the lower left of the cell. If multiple algorithms obtain the best, second best or worst value then a marker is placed for each algorithm, i.e., multiple black, gray or white markers can be placed in a row.

The markers are used to more prominently illustrate which algorithm performs the best, next best and worst on each problem. As will be shown later, the markers also provide a good means of identifying trends in algorithm performance across the set of problems.

3. ORDERED TRIAL RANK

In addition to using the proposed visualization method, a method for summarizing performance scores is proposed called the Ordered Trial Rank (OTR).

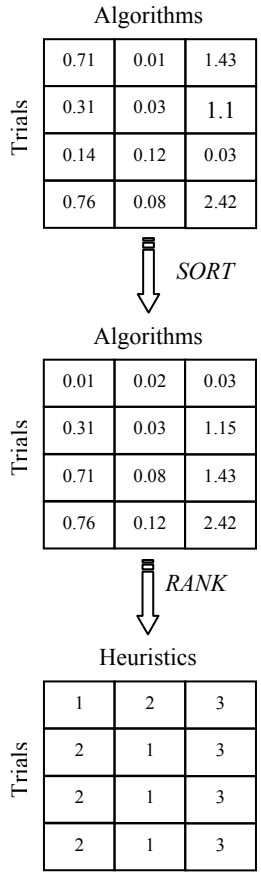

Figure 5. Illustration of the Ordered Trial Rank (OTR).

When comparing multiple algorithms on multiple problems it is often the case that the scaling of the measures is not consistent across the set of problems or indeed algorithms. For example, on a measure of objective value, some problems may produce values in the range [0, 100] while others may produce values in the range [0, 1] (or any range for that matter). It is difficult to make direct comparisons between these results without normalizing the values. However, while linearly normalizing values within the range of known values can be acceptable in some cases, it does not reflect the bias in values that can occur. A linear normalization of results assumes no or little bias in the measures of performance, which is rarely the case in real-world and difficult benchmark problems.

The benefit of the Biaxial Box Plot is that it allows for both comparisons down columns as well as across a row. For this reason it is important that the scales are appropriate and shared across all rows for the comparisons of column results to be meaningful.

One method for overcoming this problem would be to substitute the measure values for a meta-value that is more appropriately scaled. The rank of each value in the set of corresponding values would provide this meta-value with a fixed linear scale that could easily be normalized in the range [0, 1]. However, simply ranking the algorithms based on a single value, such as the average of the trial results, would again result in data loss. The Ordered Trial Rank is a method for generating appropriate rank data for collections of sets of trial results which maintains the variance results whilst providing a uniformly scaled set of values.

It should be noted that the OTR method is not applicable to paired tests, i.e., the first trial run of each algorithm uses the same seed, the second trial a different seed, and so on. However, for many optimization experiments the same seed is used for all trials. In these cases the OTR can be applied. The sets of results must be independent and ordinal, interval or ratio.

The ordered trial rank score is calculated by sorting each algorithm's trial results on each problem, shown in Figure 5. The results are then compared across all algorithms for each row of trial results and ranked, i.e., the algorithms are ranked by their best result, and then ranked by their next best result and so on. The column vector of ranks can then be averaged for each algorithm and assigned as the average ordered trial rank result for each algorithm on the given problem. The ranks can be scaled in the range [0, 1] linearly for consistency.

The sorting process ensures a fairer comparison between sets of samples rather than randomly comparing trial results in the order in which they are obtained. The sorting process ensures that the best result of each algorithm is compared with the best result from every other algorithm and, likewise, the worst results compared against the worst results.

3.1 Average Rank

While the average of the OTR score is similar to the method Average Rank, there is one important difference. OTR computes the rank of multiple samples in one dimension (which optionally can be averaged), unlike Average Rank which computes the average rank of one sample in multiple dimensions. As such, the set of average OTR results for an algorithm across multiple problems could be used to calculate an algorithms' Average Rank across the set of problems.

3.2 Statistical Testing

While the OTR measure holds similarity to the Mann–Whitney U test or Friedman test, it is important to note that it is not a statistical non-parametric rank test. Paired tests like Mann–Whitney U and Wilcoxon signed-rank are non-parametric statistical hypothesis tests which compare two related sets of samples and are used to determine whether their distributions differ. These types of tests determine whether the samples are drawn from statistically similar distributions and do not represent the distributions themselves.

Furthermore, the OTR measure can be applied to two or more related sets of samples (unlike Mann–Whitney U or Wilcoxon signed-rank) and so bares a closer resemblance to tests such as Kruskal–Wallis one-way analysis of variance by ranks or the Friedman test. The Friedman test, for example, is used to analyze two or more related samples and uses the difference in mean ranks to determine the similarity of the distributions.

While comparing the mean or median OTR values could give an indication of the statistical similarity of the samples, the OTR measure is not a statistical test in itself. Rather, the OTR provides an ordinal scale with preferable features (normalized, etc.) which are well suited to comparing sets of related distributions, such as those needed by the Biaxial Box Plot. The nature of the Biaxial Box Plot and OTR distributions allows for comparison along rows and down columns, while single statistical test values would be restricted to row comparisons.

4. ANALYSIS

An experiment was conducted demonstrate the efficacy of the proposed Biaxial Box Plot and Ordered Trial rank. A set of Evolution Strategies (ESs) were applied to a set of simple optimization test problems. Each ES was applied for 50 trial runs on each problem. Each trial was run for 5000 generations. The final best objective value obtained in each trial run was recorded. This generated a matrix of sets of trial run results for visualization.

4.1 Algorithms

A simple (1+1)-ES [8] was used as a basis for 7 optimizing algorithms. All (1+1)-ESs were given a passive archive which retained the best solutions found so far. Purely elitist selection was used. All (1+1)-ESs used only the single-point additive mutation operator. The perturbation values for the mutation operator were drawn from a Gaussian distribution. A different parameterization of the Gaussian distribution was used for each of the seven ESs, with standard deviations as follows, σ = 0.0003, 0.0015, 0.003, 0.015, 0.03, 0.15, 0.3.

4.2 Problems

The optimization problems were all variations of a simple single-parameter, single-objective problem, referred to as the *cosine problem*. The cosine problem was defined as follows:

$$f(x) = \frac{3x}{4}\left(1 - \frac{\cos(\lambda\pi x) + 1}{2}\right) + \frac{x}{4}$$

The cosine problem produced a simple landscape with deceptive optima at intervals of $\lambda\pi$, shown in Figure 6. By varying the parameter λ, different landscapes can be produced with increasing or decreasing number of local optima at varying distances, as illustrated in Figure 7.

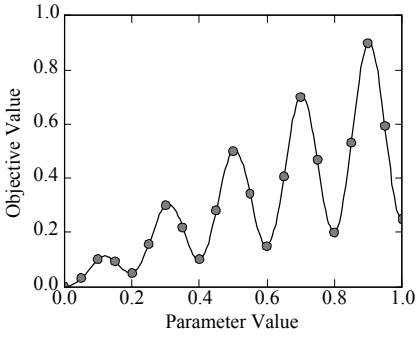

Figure 6. Illustration of the Cosine Problem.

Thirteen instances of the cosine problem were used in this study with λ ranging from 0 to 100. The first instance ($\lambda=0$) was set as a linear problem ($f(x) = x$) while all following problems were variants of the cosine problem. The full set of parameters are as follows: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50, 100.

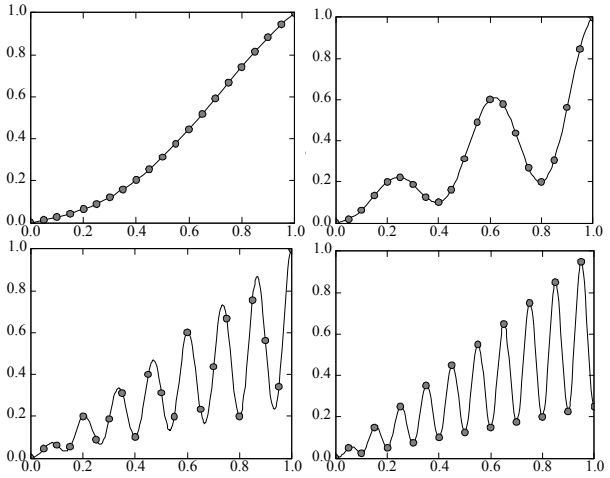

Figure 7. Demonstration of the effect of varying the cosine problem.

4.3 Results

Figure 8 shows the mean final objective value for each of the Gaussian Mutation variants of the (1+1)-ESs on the thirteen variants of the cosine test problem. The figure clearly indicates that the $\sigma = 0.3$ mutation is the most successful parameter setting across all the problems while the smallest $\sigma = 0.0003$ is the worst.

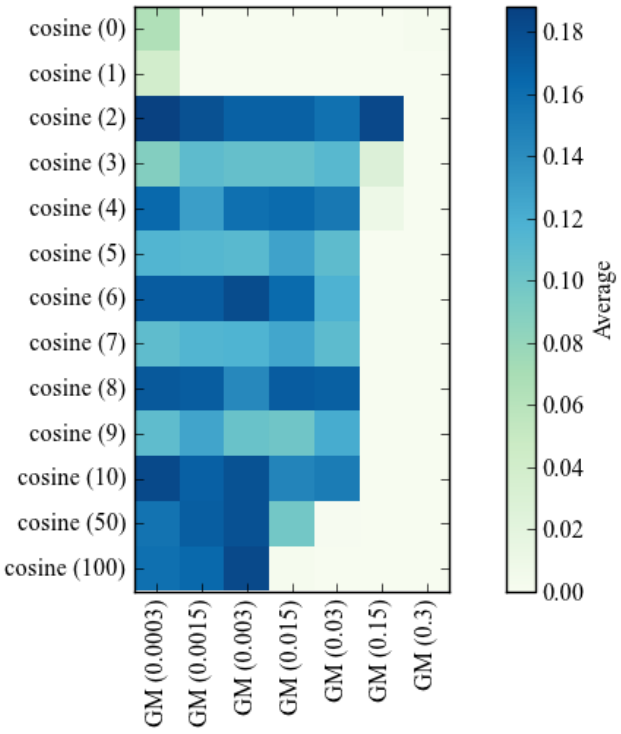

Figure 8. Heatmap showing the average best objective value for each algorithm on each problem.

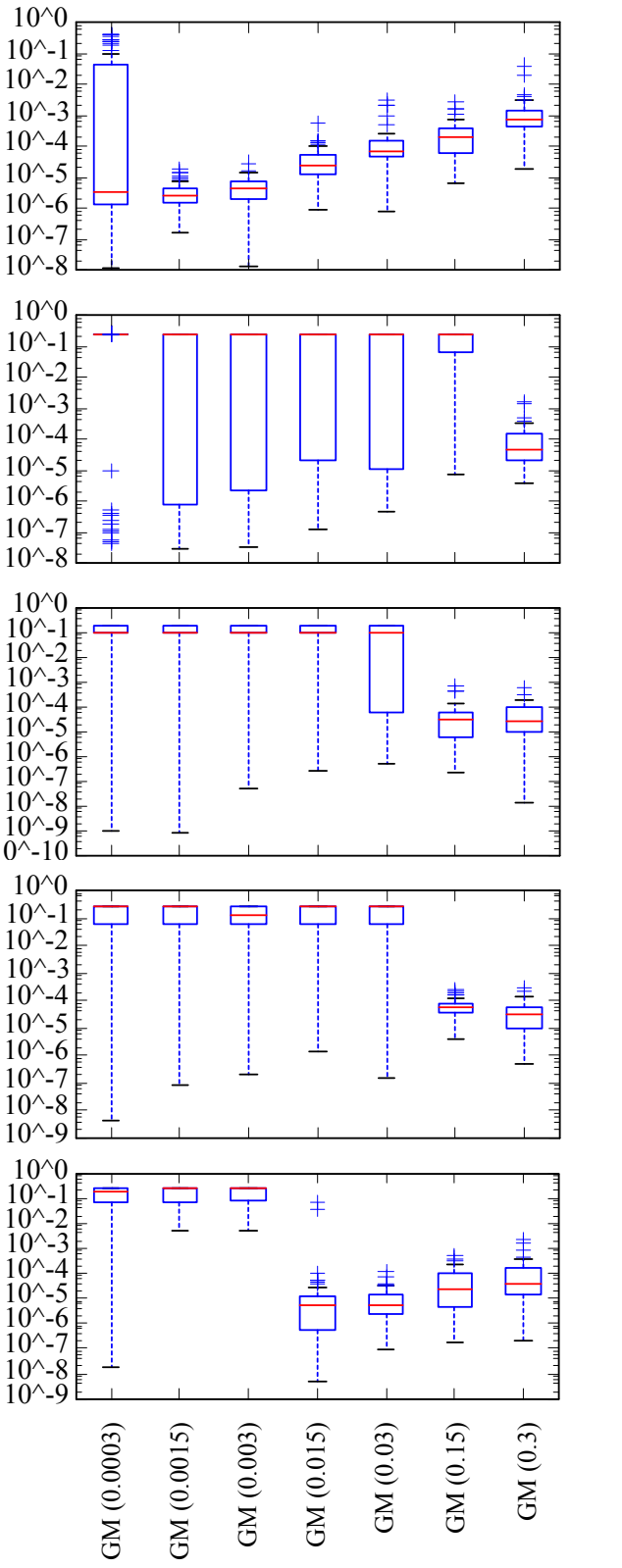

Figure 9. Box Plots showing the distribution of the Final Generation Objective Values (log scale) of each algorithm on selected problem (0, 2, 5, 8, 100)

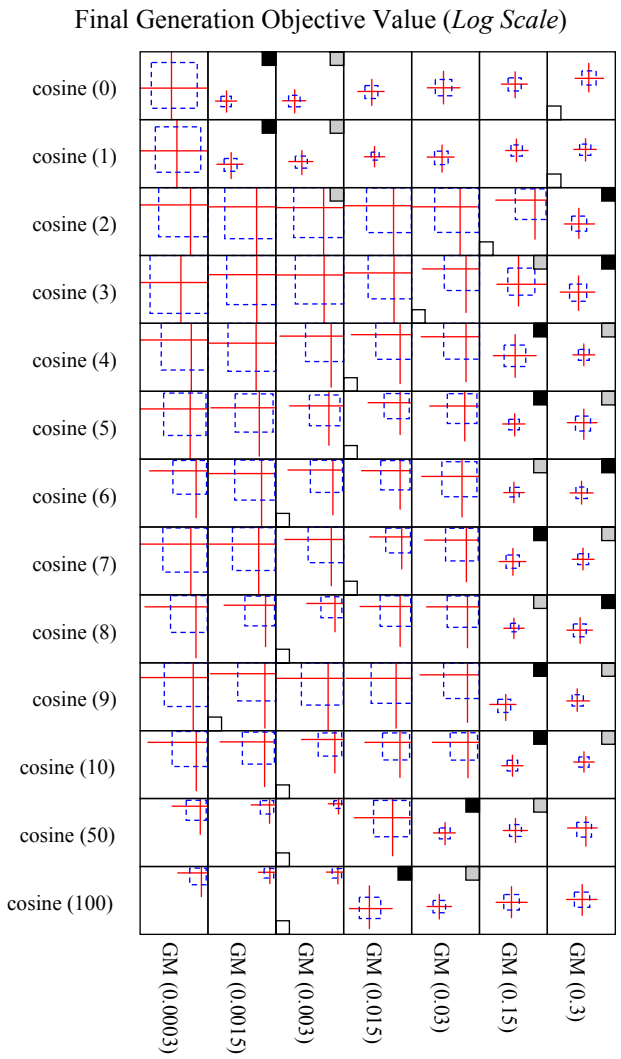

Figure 10. Biaxial Box Plot of the Final Generation Objective Values (log scale) of each algorithm on each problem.

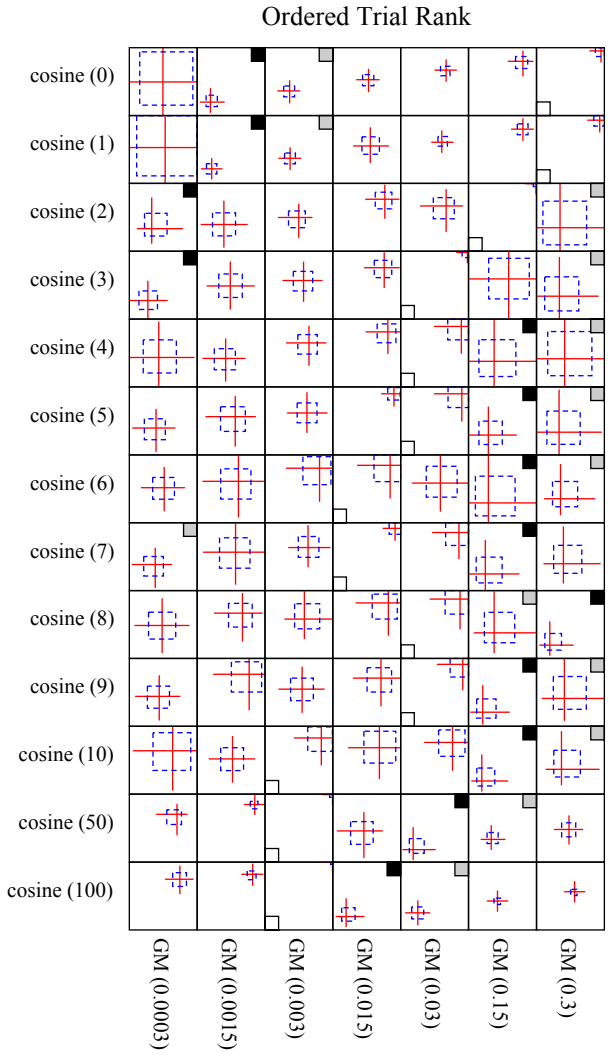

Figure 11. Biaxial Box Plot of the Ordered Trial Rank of each algorithm on each problem based on the best objective value result from each trial run.

Figure 8 displays clearly the largest differences in average values between the darker and mid blue colours and the light blue and white colours. However, in this plot it is difficult to distinguish between the lightest colours and also to determine the approximate mean objective value of each algorithm-problem pairing.

Figure 9 shows some selected problem results in more detail. This figure demonstrates how the data shown in the heatmap (Figure 8) lacks the detail given in the box plots. For example, the first box plot (*cosine (0)*) illustrates the significant difference in performance between each of the algorithms despite each having a very similar white colour in the heatmap, with the second mutation variant producing results nearly three orders of magnitude better than the last on the first problem.

In addition to providing a clearer visualization of the numeric value of the results, the box plots also supplement this with information about the distribution of results. For example, the poor mean average value of the first mutation variant shown in the heatmap can be seen to be a result of the heavy tail of distribution of results, shown in the first box plot of Figure 9. Furthermore, the results shown in the second plot (*cosine (2)*) illustrate a better distribution of results for $\sigma = 0.03$ compared to the smaller distributions whilst having similar median and mean averages – another important result which distinguishes between the algorithms' performances.

If these results were generated for the purpose of informing the selection of a suitable parameter value for solving the cosine problem, the heatmap visualization would appear to support the selection of the last parameter value while the 5 selected box plots indicate that the $\sigma = 0.15$ parameter value might equally be a good choice. If the majority of problem instances were with the lowest frequency values with a smoother landscape then the smaller distributions would also be a possible consideration.

Clearly, while the heatmap provides an easily read overview of the results, it cannot be used reliably as a good indicator of the

actual comparative results of each algorithm on each problem. Conversely, while the box plot results shown in Figure 9 give a much more detailed visualization of the results, the figures are verbose and require significantly larger page space which is a strong limiting factor in modern scientific publications. The results are less easily compared between problems in the box plot format where the y-axis scales vary between plots and their spatial arrangement is less conducive to easy comparison.

Figure 10 shows the same log scaled final generation objective values as given in Figure 9 using the Biaxial Box Plot (scaled [0, 1]). The Biaxial Box Plot shows the same median average and distribution shape results as the traditional box plots but in a more condensed matrix format which provides the same overview as the heatmap.

The black and grey markers clearly indicate that both the $\sigma = 0.15$ and $\sigma = 0.3$ mutation parameter values are effective across the whole range of problem instances. The white markers indicate that there is a shift in the worst result from the larger distributions on the smoother landscapes to the smaller distributions on the rougher landscapes.

Reading across the rows, in the same way as the box plots, it can be seen than the second smallest Gaussian distribution variant is the most effective parameter selection on the cosine (0) and cosine (1). Reading down the columns, it can be observed that the smallest $\sigma = 0.0003$ algorithm variant becomes increasing worse as the problem frequency increases. Similarly, the $\sigma = 0.15$ and $\sigma = 0.3$ algorithms are less stable on the cosine (2) to cosine (10) problems, producing larger variances in their results.

The results illustrated in Figure 11 demonstrate the efficacy of the Biaxial Box Plot for displaying large sets of experimental results. The matrix format enables easy comparison of results between algorithms and problems which is aided by the simplified, mirrored box plot format which provides similar detail on the distribution of results as given in the larger box plot format.

4.4 Ordered Trial Rank

Figure 11 uses a Biaxial Box Plot to visualize the ordered trial rank values of the experiment results. The OTR results are useful for identifying relative algorithm performance and trends in results.

The black and gray markers (best results) replicate much of the same results given in Figure 10 which uses the raw objective value results which gives confidence in the OTRs transformation of the result data.

However, there are some differences, such as the cosine (2) and cosine (3) results where the first distribution is marked as having a better average ordered trial rank. When examining the results in Figure 10, the best result given in Figure 11 is made clear by the large distribution of results produced by the smaller distribution. Clearly, when the algorithm is able to locate the region of the optimal parameter value it is able to find significantly better objective value results. Consequently, for many of the trial results, this mutation variant obtains the best result and therefore obtains a better average OTR score.

The cosine(0) results also clearly delineate the relative performance results of the algorithms, with the second mutation variant consistently outperforming the other algorithms.

The OTR results also highlight the variability in the results of the $\sigma = 0.15$ and $\sigma = 0.3$ mutation variances with large variances in their OTR scores. This is a consequence of the similar performance of these two algorithms as well as the interfering but less frequent good results generated by the other algorithms. These results suggest that the $\sigma = 0.15$ and $\sigma = 0.3$ results are not significantly different from one another.

The results given in Figure 11 demonstrate how the OTR can be used to statistically examine the relative results of many algorithms on many problems over a number of trial optimization runs and compliments the visualization of the raw data results.

5. CONCLUSION

This paper presented a novel visualization method called the Biaxial Box Plot which is designed for displaying large sets of experimental data. The Biaxial Box Plot uses the matrix format of heatmaps as a template and restructures the box plots format to show slightly less detailed information but across a wider set of data.

In addition, this paper presented a novel method for analyzing experimental data with multiple sets of samples – such as a number of algorithms' performance on an optimization problem. The method, called the Ordered Trial Rank (OTR), is easily incorporated into the Biaxial Box Plot and provides a visual means of comparing the statistical significance of individual algorithm-problem pairs as well as performance trends of different algorithms on one problem or an algorithm across multiple problems.

An illustrative experiment was conducted to produce sample optimization results of 7 algorithms on 13 problems. The results were used to demonstrate the efficacy of the Biaxial Box Plot for displaying large sets of experimental results when compared to heatmaps and traditional box plots. The matrix format enables easy comparison of results between algorithms and problems which is aided by the simplified, mirrored box plot format which provides similar detail on the distribution of results as given in the larger box plot format. The results also illustrated how the OTR can be used to statistically examine the relative results of many algorithms on many problems over a number of trial optimization runs and compliments the visualization of the raw data results.

6. REFERENCES

[1] Coello Coello, C. A., Lamont, G. B. & Van Veldhuizen, D. A., 2007. Evolutionary Algorithms for Solving Multi-Objective Problems. Second ed. New York: Springer.

[2] Deb, K., Agrawal, S., Pratab, A. & Meyarivan, T., 2000. A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. Paris, France, Springer. Lecture Notes in Computer Science No. 1917, pp. 849-858.

[3] Burke, E. et al., 2003. Hyper-heuristics: an Emerging Direction in Modern Search Technology. In: F. Glover & K. G.A., eds. Handbook of Metaheuristics. s.l.:Kluwer.

[4] Ong, Y. S., Lim, M. H., Zhu, N. & Wong, K. W., 2006. Classification of Adaptive Memetic Algorithms: A Comparative Study. IEEE Trans. on Sys., Man and Cybernetics - Part B, 36(1), pp. 141-152.

[5] Deb, K., Thiele, L., Laumanns, M. & Ztizler, E., 2001. Scalable Test Problems for Evolutionary Multi-Objective Optimization, Zurich, Switzerland.

[6] Huband, S., Hingston, P., Barone, L. & While, L., 2006. A Review of Multiobjective Test Problems and a Scalable Test Problem Toolkit. In IEEE Trans. on Evolutionary Computation, 10(5), pp. 477-506.

[7] Li, H. & Zhang, Q., 2009. Multiobjective Optimization Problems with Complicated Pareto Sets, MOEA/D and NSGA-II. IEEE Trans on Evolutionary Computation, April, 12(2), pp. 284-302.

[8] H.-G. Beyer, H.-P. Schwefel, 2002. Evolution Strategies: A Comprehensive Introduction. Journal Natural Computing, 1(1):3–52.

[9] P. J. Rousseeuw, I. Ruts, J. W. Tukey, 1999. The Bagplot: A bivariate boxplot, American Statistician 53(4): 382-38.