A Large-Scale Experimental Evaluation of High-Performing Multi- and Many-Objective Evolutionary Algorithms

Scenario-wise Analysis

Leonardo C. T. Bezerra leo.tbezerra@ci.ufpb.br
DCC, Universidade Federal da Paraíba, João Pessoa, PB, Brazil

Manuel López-Ibáñez manuel.lopez-ibanez@manchester.ac.uk
Alliance Manchester Business School, University of Manchester, UK

Thomas Stützle stuetzle@ulb.ac.be
IRIDIA, CoDE, Université Libre de Bruxelles, Belgium

1 In-depth analysis

In this section, we discuss results grouped first by the number of objectives, making comments on the overall results by means of a rank sum analysis. We then proceed to boxplot analysis of all metrics grouped by the number of FEs given to MOEAs.

1.1 Two-objective problems

Results for two-objective problems are given in Table 1. For each row, we sort algorithms according to their rank sums when ran for the given number of FEs and assessed by the given performance metric. Algorithms highlighted (**boldface**) present rank sums considered statistically significantly lower than the others according to Friedman's test with 99% confidence level. As we will shortly discuss in more detail, we can observe three groups of algorithms according to their performance throughout scenarios. In general, SMS and IBEA are the algorithms that present best performance. The second group of algorithms comprises NSGA-II, SPEA2, HypE, MOEA/D and NSGA-III, which also present overall good performance, but are affected by specific function characteristics or by the FE budget they are given. Finally, the last group is formed by MO-CMA-ES and MOGA. The former often presents poor performance, in part due to its necessity of large FE budgets; the latter is consistently worse than almost all other MOEAs, both due to its lack of elitism and its lack of both limit-stability and limit-optimality.

Table 1: Sum of ranks (in parenthesis) depicting the performance of MOEAs on two-objective problems.

2500 FEs									
I_H^{rpd}	**SMS** (0)	**IBEA** (11)	NSGA-II (456)	SPEA2 (971)	MOEA/D (1972)	HypE (2124)	NSGA-III (2396)	CMA (5139)	MOGA (5622)
I_ϵ	**SMS** (0)	**IBEA** (159)	NSGA-II (1806)	SPEA2 (2048)	HypE (2071)	NSGA-III (2446)	MOEA/D (3976)	CMA (5766)	MOGA (6017)
IGD	**SMS** (0)	IBEA (490)	NSGA-III (2979)	SPEA2 (3025)	HypE (3152)	NSGA-II (3172)	CMA (5494)	MOEA/D (6026)	MOGA (6386)
10000 FEs									
I_H^{rpd}	**IBEA** (0)	**SMS** (113)	SPEA2 (1149)	NSGA-II (1846)	MOEA/D (2819)	HypE (3593)	CMA (3731)	NSGA-III (4017)	MOGA (6682)
I_ϵ	**SMS** (0)	IBEA (425)	SPEA2 (894)	NSGA-II (1993)	MOEA/D (3091)	HypE (3401)	CMA (3749)	NSGA-III (4231)	MOGA (6740)
IGD	**SMS** (0)	SPEA2 (711)	IBEA (1387)	HypE (2382)	NSGA-II (2774)	MOEA/D (4237)	NSGA-III (5068)	CMA (5240)	MOGA (7297)
40000 FEs									
I_H^{rpd}	**SMS** (0)	IBEA (453)	SPEA2 (791)	NSGA-II (1795)	MOEA/D (2269)	NSGA-III (3492)	CMA (3818)	HypE (3888)	MOGA (6788)
I_ϵ	**SMS** (0)	SPEA2 (918)	IBEA (1197)	NSGA-II (2613)	MOEA/D (3053)	NSGA-III (3208)	CMA (3393)	HypE (4295)	MOGA (7060)
IGD	**SMS** (0)	SPEA2 (731)	IBEA (2404)	MOEA/D (2833)	NSGA-III (3039)	NSGA-II (3409)	HypE (4252)	CMA (4446)	MOGA (7364)

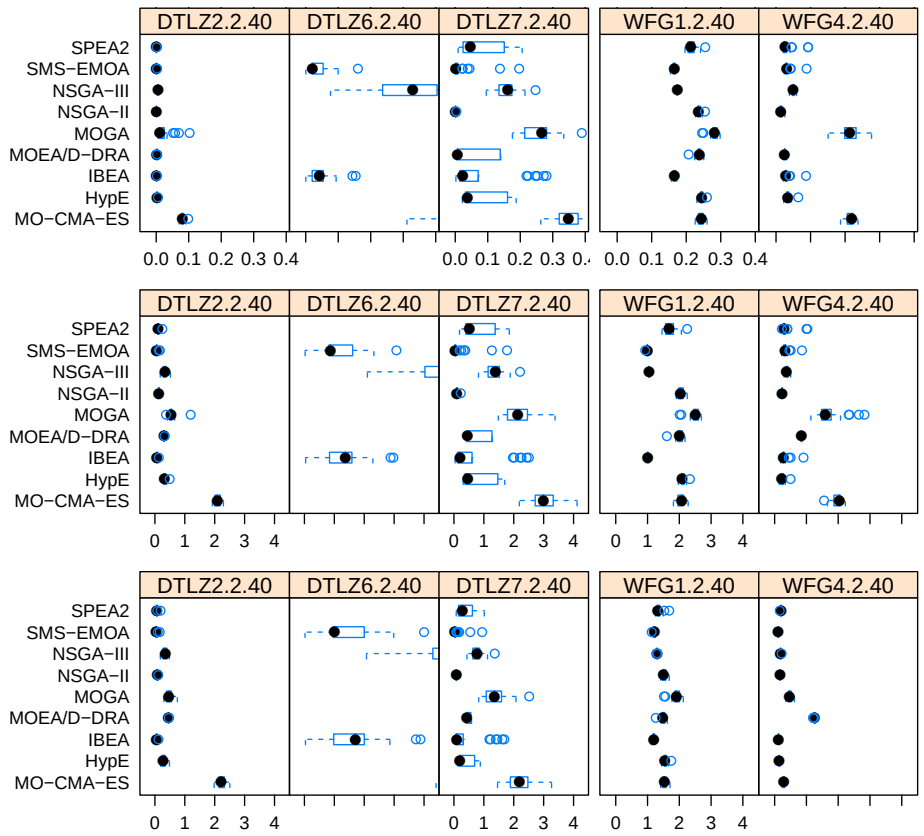

Figure 1: Performances of MOEAs given 2500 FEs on selected two-objective problems with 40 variables. From top to bottom, I_H^{rpd}, I_ϵ and *IGD*.

We then proceed to a more detailed discussion grouping results by the number of function evaluations.

1.1.1 2500 FEs

Figure 1 shows the performance of all MOEAs on selected problems when given 2 500 FEs. For brevity, we focus the discussion on the plots that better illustrate the rankings given in Table 1. Moreover, the benchmark problems we consider here can be grouped according to similar characteristics or difficulty they pose for MOEAs, and the problems we depict in Figure 1 are representative of each of these groups, as we will detail.

Given the low budget of FEs, it is not surprising that many MOEAs are unable to converge to the optimal front. In general, MOEAs display good performance on the easiest DTLZ problems, namely DTLZ2 and DTLZ5. The exception to this pattern are the worst-ranked MOEAs, i.e., MOGA, MO-CMA-ES, and MOEA/D. In fact, we remark that for this couple problems MO-CMA-ES ranks far worse than all other MOEAs according to all metrics. In addition, while MOEA/D performs slightly better than MOGA according to the I_H^{rpd} and the $I_{\epsilon+}^1$, it is outperformed according to the *IGD*. Another important observation is that, except for MO-CMA-ES, the I_H^{rpd} performance of all MOEAs look fairly similar, with differences being noticed mostly on the remaining indicators.

By contrast to the good performance displayed by nearly all MOEAs on the easiest DTLZ problems, their performance on the hardest problems (DTLZ1, DTLZ3, and DTLZ6) is astonishingly poor. In fact, DTLZ6 is the only problem where a few MOEAs are still able to get close to the actual front. The number of local optimal fronts make

these problems too difficult for MOEAs when given a low budget of FEs, as we have discussed in the paper. The only algorithms that are able to display reasonable performance are the best-ranked ones, namely SMS and IBEA.

The moderately difficult DTLZ problems are the ones that present bias (DTLZ4) or a multi-modal, disconnected front (DTLZ7). The performance patterns on each of these problems is unique, so we address them individually. For DTLZ4, most MOEAs get trapped in biased regions of the search space, with SMS and IBEA being the only algorithms that are able to properly approximate the front on the majority of their runs. In addition, we remark that the performance of MO-CMA-ES according to the distance-based metrics is very poor. Concerning DTLZ7, SMS and NSGA-III are the algorithms that display best performance, being able to accurately approximate the Pareto optimal front on most their runs. In addition, MOEA/D is also able to display considerably good performance according to the I_H^{rpd}, but the distance-based metrics disagree. Again, MO-CMA-ES presents rather poor performance, being far worse than MOGA.

As for the WFG problems, these can be grouped into two major difficulty levels. The first, and harder, is formed by the two convex problems WFG1–2. As one can see from the plots for WFG1, no MOEA is able to reach the actual front. Nonetheless, SMS is the best-performing algorithm for this group of problems, followed by IBEA on WFG1 and by NSGA-III on WFG2. We also remark that the performance metrics disagree on these convex problems: while the I_H^{rpd} and the $I_{\epsilon+}^1$ show a large performance difference between MOEAs, the *IGD* shows a much narrower scenario. Nevertheless, all metrics agree that much improvement could be achieved by MOEAs in general.

The second group of WFG problems presents moderate difficulty for a few MOEAs. Although none is able to reach the front with the reduced number of FEs, results presented by MOEA/D, MO-CMA-ES, and MOGA are far worse than the ones presented by the remaining MOEAs. Once again, we see a considerable performance difference for a few algorithms depending on the metric considered, namely MO-CMA-ES and MOEA/D. For the I_H^{rpd} and the $I_{\epsilon+}^1$, the performance of MOEA/D is not very worse than most MOEAs, while MO-CMA-ES is the worst-ranking algorithm. By contrast, the *IGD* indicates the opposite.

1.1.2 10 000 FEs

The performance of all MOEAs when given 10 000 FEs is shown on Fig. 2 (top). In particular, since the plots for all metrics are very similar, we only depict the results for the I_H^{rpd}. The first important difference we notice concerns the group of easy DTLZ problems, represented by DTLZ2. More specifically, the only MOEA that is still unable to reach the optimal front is MOGA. Concerning the hardest problems, the performance of all MOEAs is now much better on DTLZ6, but DTLZ1 and DTLZ3 remain unfeasible for all MOEAs. We also remark that the performance of MO-CMA-ES and the decomposition-based algorithms are particularly affected by the characteristics of this problem. Although the three MOEAs fail to correctly approximate the front, NSGA-III shows much worse performance than the other two, and MOEA/D is outperformed by a large margin by MO-CMA-ES. Finally, for the moderately difficult DTLZ problems (DTLZ4 and DTLZ7), most MOEAs are able to reach the optimal front, except for MOGA and MOEA/D on both problems, NSGA-III on DTLZ4, and MO-CMA-ES on DTLZ7.

Figure 2: Performances of MOEAs on selected two-objective problems with 40 variables. Since all metrics agree, we display only the I_H^{rpd}. On top, performance for 10 000 FEs. On the bottom, performance for 40 000 FEs.

When we analyze the performance of the algorithms on the WFG problems, we see that all algorithms benefit from the extra FEs on the hardest functions (WFG1–2). However, while SMS and HypE are able to correctly approximate the optimal front of WFG2, no MOEA is able to so on WFG1. Similarly to the previous scenario, SMS is the best-performing algorithm on this couple problems. Once again, NSGA-III is outperformed by all MOEAs but MOGA, indicating that this algorithms lacks robustness w.r.t. different problem characteristics. As for the remaining WFG functions, the performance of all MOEAs is greatly improved, with most of the indicator- and dominance-based algorithms performing nearly equivalently. The only exception to this pattern is MOGA, which still performs poorly.

1.1.3 40 000 FEs

The substantial increase in the number of function evaluations given to MOEAs reflects in performance improvements from most MOEAs which had failed to converge in previous scenarios, as we can see in Fig. 2. On the easiest DTLZ, all MOEAs but MOGA converge to the actual fronts except for MOGA. On the moderate, MOEA/D and NSGA-III still face difficulties on DTLZ4, whereas MO-CMA-ES struggles on DTLZ7. On the hardest, no MOEA is able to reach the optimal front for DTLZ1 and DTLZ3, indicating that these problems are probably unfeasible for MOEAs when moderate number of variables are used. Another important observation is that even with this increased FE budget, NSGA-III fails to accurately approximate the Pareto optimal front for DTLZ6. The same happens to MO-CMA-ES and MOEA/D on the larger n_{var} values.

Concerning WFG functions, most MOEAs are able to converge to the actual fronts on the WFG3–WFG9 functions, MOGA being the exception. The same performance improvements can be observed for WFG1, with SMS, IBEA, and SPEA2 reaching excellent results. By contrast, a lot of variability can be seen on the results of the best-performing MOEAs on WFG2 according to both the I_H^{rpd} and the $I_{\epsilon+}^1$, HypE being the exception. Finally, the decomposition-based and MO-CMA-ES display the worst performance on both these problems.

Table 2: Sum of ranks (in parenthesis) depicting the performance of MOEAs on three-objective problems.

2500 FEs									
I_H^{rpd}	**SMS** (0)	IBEA (743)	HypE (1516)	MOEA/D (2421)	SPEA2 (3464)	NSGA-II (3950)	NSGA-III (4068)	CMA (4928)	MOGA (6727)
I_ϵ	**SMS** (0)	**IBEA** (53)	MOEA/D (2456)	HypE (3102)	NSGA-II (3420)	NSGA-III (3580)	SPEA2 (4421)	CMA (5495)	MOGA (6352)
IGD	**IBEA** (0)	MOEA/D (650)	SMS (711)	NSGA-II (1434)	NSGA-III (2710)	SPEA2 (3592)	HypE (3944)	CMA (4883)	MOGA (5420)
10000 FEs									
I_H^{rpd}	**SMS** (0)	IBEA (556)	MOEA/D (1805)	HypE (2290)	SPEA2 (2302)	CMA (3616)	NSGA-II (4378)	NSGA-III (4627)	MOGA (7029)
I_ϵ	**SMS** (0)	IBEA (494)	SPEA2 (2516)	CMA (2968)	HypE (3132)	MOEA/D (3552)	NSGA-III (4253)	NSGA-II (4885)	MOGA (7323)
IGD	**IBEA** (0)	SMS (654)	SPEA2 (1041)	MOEA/D (1622)	HypE (2960)	NSGA-II (3640)	CMA (4240)	NSGA-III (4264)	MOGA (6886)
40000 FEs									
I_H^{rpd}	**SMS** (0)	IBEA (715)	MOEA/D (1575)	SPEA2 (2715)	HypE (3510)	CMA (3604)	NSGA-II (4269)	NSGA-III (4276)	MOGA (7163)
I_ϵ	**SMS** (0)	IBEA (994)	CMA (2614)	SPEA2 (2708)	MOEA/D (3137)	NSGA-II (4450)	NSGA-III (4774)	HypE (5284)	MOGA (7610)
IGD	**MOEA/D** (0)	SMS (269)	SPEA2 (763)	IBEA (1668)	NSGA-II (2780)	CMA (3118)	NSGA-III (3984)	HypE (4815)	MOGA (6686)

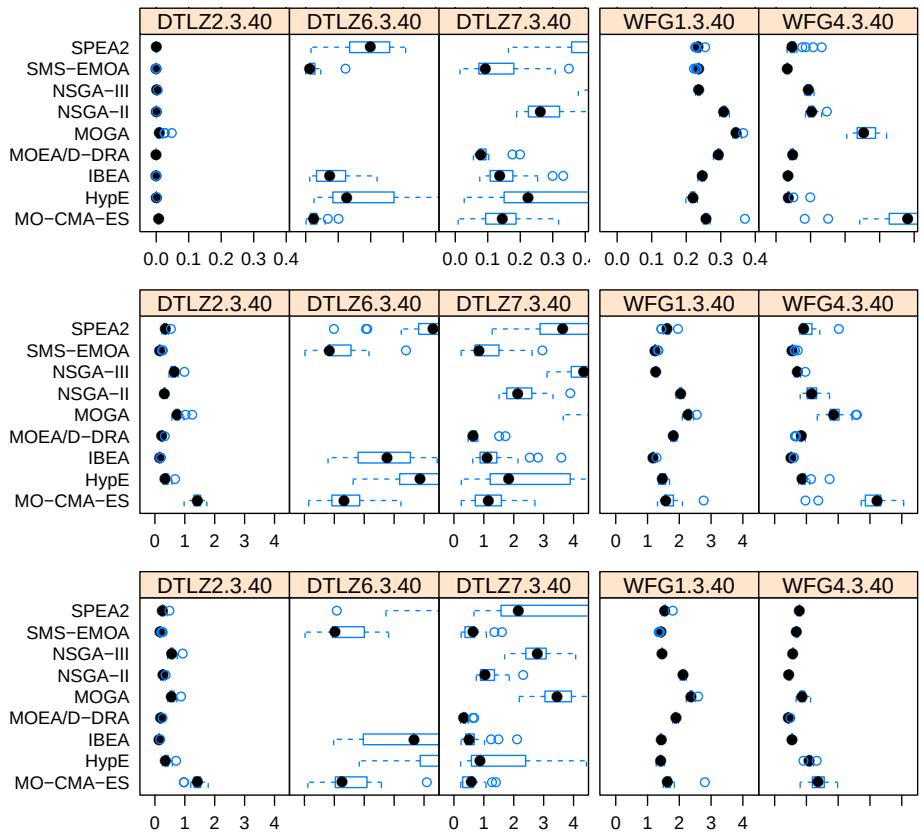

Figure 3: Performances of MOEAs given 2 500 FEs on selected three-objective problems with 40 variables. From top to bottom, I_H^{rpd}, I_ϵ and IGD.

1.2 Three-objective problems

The rank sum analysis of the performance presented by all MOEAs on three-objective problems is given in Table 2. In general, the same patterns observed for the two-objective problems can be seen in this case, i.e., algorithms can be clustered into three different groups according to performance. Once again, SMS and IBEA comprise the best-performing group, but this time only MOGA comprises the worst-performing one. In addition, overall performances of HypE, MOEA/D and MO-CMA-ES are greatly improved. In the case of the MO-CMA-ES, we notice that both this algorithm and SPEA2 display much worse performance when given a low FE budget. The opposite happens to HypE, as we discuss in more detail next.

1.2.1 2 500 FEs

Figure 3 shows the performance of all MOEAs on selected problems when given 2500 FEs. For the easiest DTLZ problems (DTLZ2 and DTLZ5), MOEAs are able to display good performance according to the I_H^{rpd}, but the distance-based metrics favor SMS and IBEA over all other algorithms. In particular, MO-CMA-ES is the worst-performing algorithm, although we remark NSGA-III and performs as poorly as MOGA.

The increase in the number of objectives poses more difficulty for MOEAs also for the moderate DTLZ functions (DTLZ4 and DTLZ7). Once again, SMS and IBEA are the best-performing algorithms, although we see that, particularly for DTLZ7, SPEA2 faces a significant challenge, whereas MO-CMA-ES appears to deal with its characteristics quite favorably. In fact, this algorithm outperforms all other MOEAs on DTLZ7 when $n_{\text{var}} = 30$, but is not able to maintain its top-performing behavior on larger n_{var}

values. Concerning metrics, two very different patterns are observed. For DTLZ4, the distance-based metrics agree and show that many MOEAs have much room to improve, whereas the I_H^{rpd} would seem to indicate that all algorithms have successfully converged. By contrast, on DTLZ7 the opposite happens, with I_H^{rpd} results being worse than the distance-based metrics.

For the hardest DTLZ problems, two different situations are observed. For DTLZ1 and DTLZ3, again no MOEA is able to approximate the fronts. For DTLZ6, however, the performance of some MOEAs is actually better than for the two-objective DTLZ6, in particular HypE, MO-CMA-ES, and SPEA2. In fact, MO-CMA-ES is only outperformed by SMS regardless of n_{var}.

Concerning the WFG problems, the group comprising the hardest ones (WFG1 and WFG2) pose a challenge similar to that observed on $M = 2$ scenarios. No algorithm is able present reasonable results whatever the performance metric considered on both problems, although some algorithms clearly perform better than others. The only exception to this pattern are SMS and IBEA, which perform well according to all metrics on WFG2 with $n_{\text{var}} = 30$, and HypE which performs well on the same problem but only according to the I_H^{rpd}. For the remaining WFG problems (represented by WFG4) we notice that algorithms are unable to reach the Pareto optimal front, although results can be considered reasonable given the limited FE budget MOEAs are allowed to use. Nonetheless, the only algorithm that is not affected by specific problem characteristics from this WFG subset is IBEA. In fact, the concave WFG problems are directly responsible for the good ranking of MOEA/D according to the *IGD*. In addition, the same indicator ranks SMS as the third-best MOEA, a significant difference w.r.t. the rest of the metrics. This fact is explained by the performance of SMS on the problems that present parameter-dependent bias, i.e., WFG7–9, where this MOEA performs slightly worse than the top two ranked algorithms.

1.2.2 10 000 FEs

Boxplots depicting the performance of MOEAs when ran for 10 000 FEs are given in Fig. 4. We initially make a few remarks concerning performance metrics. On the DTLZ benchmark, all metrics agree, although two important exceptions can be observed. Firstly, on DTLZ6 the I_H^{rpd} agrees with the other metrics, although it seems to indicate a better performance from MOEAs in general than the remaining metrics. The same pattern is observed on DTLZ7, but now it is the *IGD* that indicates a better performance than the other metrics. This also happens on WFG2, the other disconnected problem, and on the concave WFG problems. On the remaining convex WFG problems (WFG1 and WFG3), all metrics agree.

In general, the approximation fronts produced by MOEAs now present much better performance, whatever the metric considered. For the easiest DTLZ problems, all MOEAs are now able to accurately approximate the fronts. On the moderate ones, we see two different situations. On DTLZ4, improvements are more clear on the distance-based metrics, whereas I_H^{rpd} improvements are mostly seen when $n_{\text{var}} = 50$. Conversely, on DTLZ7 many MOEAs display major performance gains. The most significant exception is SMS, which is not able to improve over its performance when given $FE_{\text{max}} = 2\,500$. In fact, on this particular problem SMS would outrank only MOEA/D, which presents great variability, and MOGA.

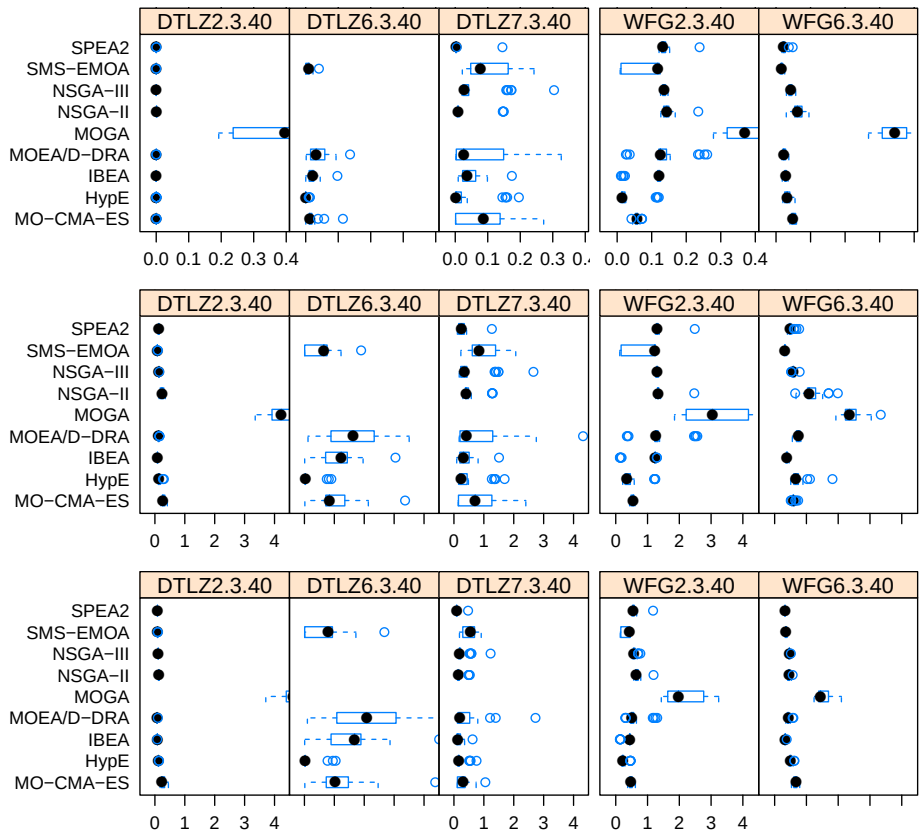

Figure 4: Performances of MOEAs given 10 000 FEs on selected three-objective problems with 40 variables. Since most MOEAs are only able to improve their I_H^{rpd} performance, we show only the plots for this metric.

For the hardest DTLZ functions, over half of the MOEAs is able to present reasonable results on DTLZ6, but none on DTLZ1 and DTLZ3. However, even on DTLZ6 we remark that the MOEAs that display very good performance on smaller n_{var} values face much more difficulties when this factor is increased. Concerning WFG problems, all algorithms show performance improvements on all functions, but none is able to reach the actual fronts. The major exception is WFG2, where some MOEAs improve their performances on extreme n_{var} values at the cost of a worsened performance on $n_{\text{var}} = 40$.

1.2.3 40 000 FEs

The increase in the computational budget reflects in different performance improvement rates that vary according to the group of problems considered. On both the easiest and the moderate DTLZ problems, most algorithms are able to properly approximate the optimal fronts. The most surprising exception is SMS, which fails to converge on DTLZ7 even when given this increased FE budget. On the hardest WFG problems, algorithms are also able to display improved performances, although on the WFG2 problem this is only observed more clearly on specific n_{var} values. Nonetheless, the only MOEAs that correctly approximate the optimal front on this problem are SMS and IBEA, and none is able to do so on WFG1. Finally, on the concave WFG problems all MOEAs show improvements, in particular according to the distance-based metrics, and even more so according to the *IGD*. Nonetheless, MOEAs are unable to converge to the actual fronts, even though some algorithms get very close to that goal. Finally, we remark that the lower relative performance of IBEA according to the *IGD* is explained by the improvements of other MOEAs on the WFG concave functions rather than by a

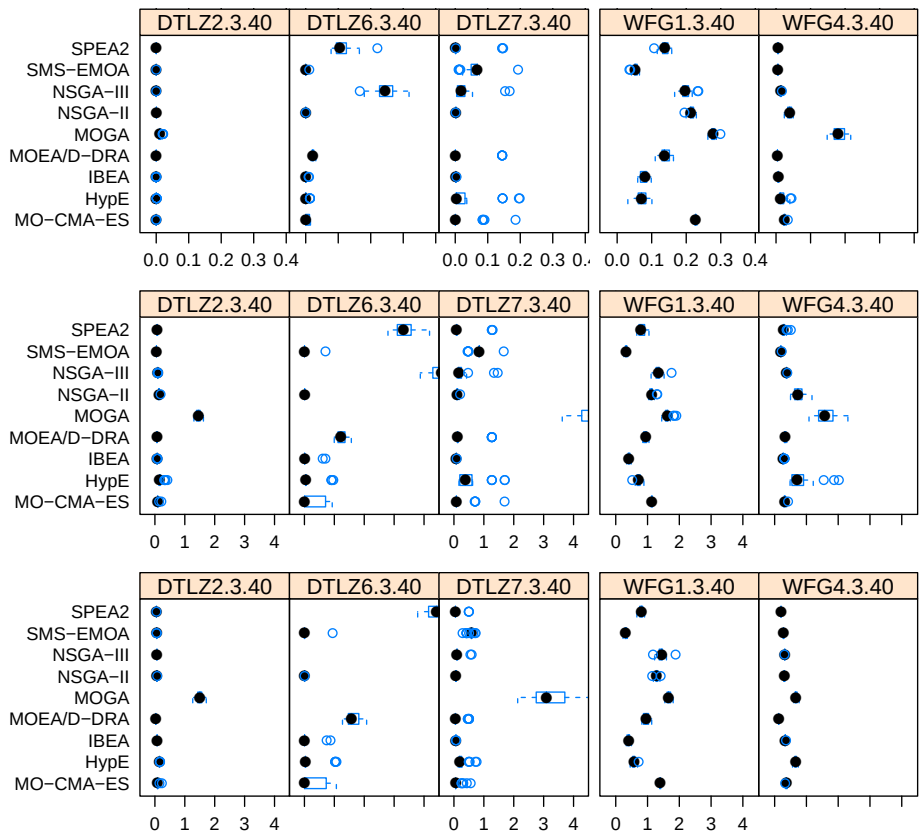

Figure 5: Performances of MOEAs given 40 000 FEs on selected three-objective problems with 40 variables. From top to bottom, I_H^{rpd}, I_ϵ and IGD.

worsening in the performance of IBEA.

Table 3: Sum of ranks (in parenthesis) depicting the performance of MOEAs on five-objective problems.

2500 FEs									
I_H^{rpd}	**SMS** (0)	IBEA (1402)	MOEA/D (1603)	SPEA2 (3375)	NSGA-II (4245)	CMA (4274)	NSGA-III (4731)	HypE (4765)	MOGA (7463)
I_ϵ	**SMS** (0)	IBEA (745)	MOEA/D (1221)	NSGA-III (2898)	NSGA-II (3766)	SPEA2 (3785)	HypE (4108)	CMA (4365)	MOGA (5588)
IGD	**SMS** (0)	**NSGA-III** (57)	**IBEA** (242)	MOEA/D (610)	HypE (1034)	SPEA2 (1052)	NSGA-II (1931)	CMA (3214)	MOGA (4503)
10000 FEs									
I_H^{rpd}	**SMS** (0)	MOEA/D (1471)	IBEA (1535)	SPEA2 (3295)	CMA (3374)	NSGA-III (3897)	NSGA-II (4130)	HypE (5811)	MOGA (7562)
I_ϵ	**SMS** (0)	IBEA (1713)	MOEA/D (2569)	CMA (2588)	NSGA-II (4086)	NSGA-III (4124)	SPEA2 (4701)	HypE (5907)	MOGA (7664)
IGD	**SMS** (0)	IBEA (1898)	MOEA/D (2119)	NSGA-II (2329)	CMA (2515)	SPEA2 (3579)	HypE (5040)	NSGA-III (5225)	MOGA (7398)
40000 FEs									
I_H^{rpd}	**SMS** (0)	IBEA (1154)	MOEA/D (1761)	CMA (2915)	SPEA2 (2995)	NSGA-III (3325)	NSGA-II (4188)	HypE (5826)	MOGA (7363)
I_ϵ	**SMS** (0)	IBEA (173)	CMA (1610)	MOEA/D (1650)	SPEA2 (3310)	NSGA-II (3988)	NSGA-III (4436)	HypE (5777)	MOGA (7097)
IGD	**IBEA** (0)	MOEA/D (244)	SMS (827)	SPEA2 (1720)	CMA (2074)	NSGA-II (2737)	NSGA-III (5087)	HypE (5189)	MOGA (6780)

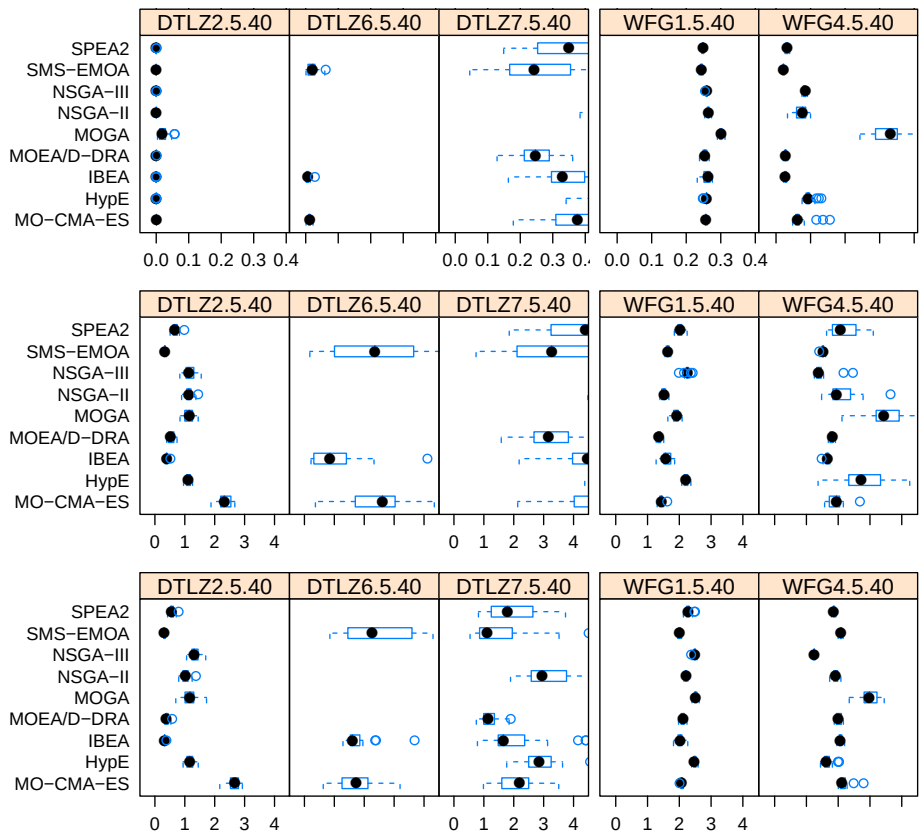

Figure 6: Performances of MOEAs given 2 500 FEs on selected five-objective problems with 40 variables. From top to bottom, I_H^{rpd}, I_ϵ and *IGD*.

1.3 Five-objective problems

As discussed in the paper, the increase in the difficulty level of the problems is much more substantial for some problems than for others. For the DTLZ problems, we discard DTLZ1 and DTLZ3 from our analysis, as no MOEA is able to produce reasonable results for any of them. We then re-categorize the remaining DTLZ problems into two groups: (i) the hardest, comprising DTLZ6 and DTLZ7, which now becomes more difficult than DTLZ6, and; (ii) the moderate, comprising DTLZ4 and the formerly easy DTLZ2 and DTLZ5. Concerning the WFG benchmark, regardless of the number of function evaluations, the performance metric and the MOEA considered, we notice that the convex problems, which before posed significant difficulty for all algorithms, now become easier to solve than the concave ones.

Initial considerations aside, the rank sum analysis of the performance presented by all MOEAs on five-objective problems is given in Table 3. Based on their performance, MOEAs can be clustered into nearly the same three groups we highlighted for three-objective problems. The major exception concerns HypE, which is now often the highest-ranked of the MOEAs that comprise the intermediate-performance group. In addition, we notice that the rankings of NSGA-III vary considerably depending on the metric considered on $FE_{\max} = 2\,500$ scenarios. We then proceed to a detailed discussion grouped by FE budget.

1.3.1 2 500 FEs

Boxplots depicting the performance of all MOEAs on three-objective problems when given 2 500 FEs are shown in Fig. 6. From the distance-related metrics one can con-

firm what we previously discussed about DTLZ2, as clearly this problem becomes a challenge for most MOEAs when a limited number of FEs is allowed. In fact, the only algorithms to converge to the actual front of DTLZ2 are SMS and IBEA. We remark that this problem is a clear example of the different metrics behavior, as the I_H^{rpd} would make it seem that this is an easy problem for all MOEAs. The performance displayed by MOEAs on the other moderate DTLZ problems (DTLZ4–5) is very similar to the one depicted for DTLZ2, confirming that the number of objectives affects the difficulty level of the problems in different degrees. In general, the other MOEA that performs nearly equivalently to the two best is MOEA/D. The most important exception concerns NSGA-III, which ranks first according to the *IGD* metric on DTLZ5.

For the hardest DTLZ problems (DTLZ6–7), we see two similar, yet slightly different situations. In common, no MOEA is able to converge to the actual Pareto front with the limited number of FEs they are given. However, while SMS, IBEA, and MO-CMA-ES are the algorithms that perform best on DTLZ6, MOEA/D is best-performing algorithm on DTLZ7 when all metrics are considered. Nonetheless, we make a few remarks concerning SMS. First, on DTLZ6 its performance is very affected by the increase in n_{var}, and so is the performance of MO-CMA-ES. Second, while the I_H^{rpd} indicates that the performances of SMS and MOEA/D are similar on DTLZ7, the distance-based metrics show a big gap between these two algorithms.

Concerning the WFG benchmark, we start our analysis with the non-concave problems, i.e., the convex WFG1–2 and the mixed linear-convex WFG3. In general, no MOEA is able to correctly approximate the optimal fronts and that it is difficult to extract patterns from those plots, but we focus on two important observations. First, on WFG2–3 SMS performs quite poorly according to the I_H^{rpd}, but much better according to the distance-based metrics. Second, although many MOEAs perform similarly across these functions, the overall best-performing MOEAs are SMS, IBEA, and MOEA/D. As for the concave WFG problems, the patterns in the boxplots depend considerably on the given metric. According to the I_H^{rpd}, NSGA-III shows the best performance, although many MOEAs display similar performance when $n_{\text{var}} = 50$. By contrast, the distance-based metrics favors SMS, followed by IBEA and NSGA-III on the $I_{\epsilon+}^1$, and by IBEA, MOEA/D, and SPEA2 on the *IGD*.

1.3.2 10 000 FEs

The increase in the number of function evaluations given to MOEAs reflects on performance improvements in nearly all problems and according to all metrics, as shown in Fig. 7. For all problems considered, no MOEA is able to properly approximate the optimal fronts. Concerning the DTLZ functions, we notice that the same group of MOEAs is able to display the best performance among all algorithms on problems DTLZ2 and DTLZ4–6 when all metrics and n_{var} values are considered altogether, namely SMS, IBEA, MOEA/D and MO-CMA-ES. By contrast, on DTLZ7 most algorithms perform similarly according to the *IGD*, but the remaining metrics favor MOEA/D considerably over the other MOEAs.

Concerning the non-concave WFG problems represented in Fig. 7 by WFG1, we see different performance patterns. For WFG1 and WFG3, all metrics agree and indicate that SMS is the best-performing algorithm, followed by SPEA2 on WFG1 and by MOEA/D on WFG3. By contrast, performance differences on WFG2 are very clear

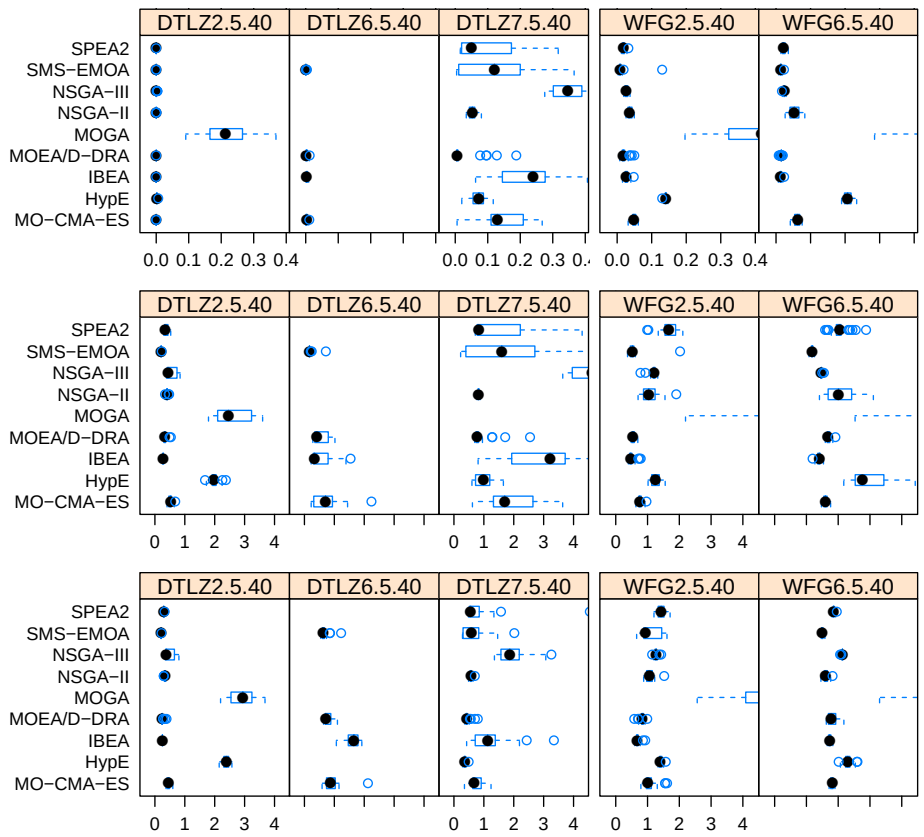

Figure 7: Performances of MOEAs given 10 000 FEs on selected five-objective problems with 40 variables. From top to bottom, I_H^{rpd}, I_ϵ and *IGD*.

when one compares results on $n_{\text{var}} = 30$ and $n_{\text{var}} = 50$. Overall, SMS, IBEA, and MOEA/D can be considered the best-performing algorithms for this function, but the performance of SMS according to the *IGD* is much worse than for the other metrics. As for the concave WFG problems, we see again these three algorithms being considered best, but we remark that NSGA-III would have made it into that group if not for its performance according to the *IGD*.

1.3.3 40 000 FEs

Boxplots depicting MOEA performances when given 40 000 FEs are shown in Fig. 8. Overall, the performance gains are clear for all MOEAs according to all metrics, except for the concave WFG problems where no significant I_H^{rpd} improvements can be seen. Nonetheless, once again no MOEA is able to successfully approximate the actual fronts, even though many MOEAs get very close to it on DTLZ2. Concerning the moderate DTLZ problems, the best-performing MOEAs get reasonably close to the actual fronts, namely SMS, IBEA, MOEA/D, and MO-CMA-ES, although the latter two lose performance on DTLZ4 when n_{var} is increased. These same four MOEAs repeats their good performance on DTLZ6. However, while other MOEAs were able to show reasonable results on the moderate problems, this time the performance of the remaining algorithms is quite poor. Finally, a peculiar result can be observed on DTLZ7: algorithms that displayed good performance on most other problems, namely SMS, IBEA, MOEA/D, and NSGA-III, are outranked by other MOEAs such as SPEA2, NSGA-II, and HypE.

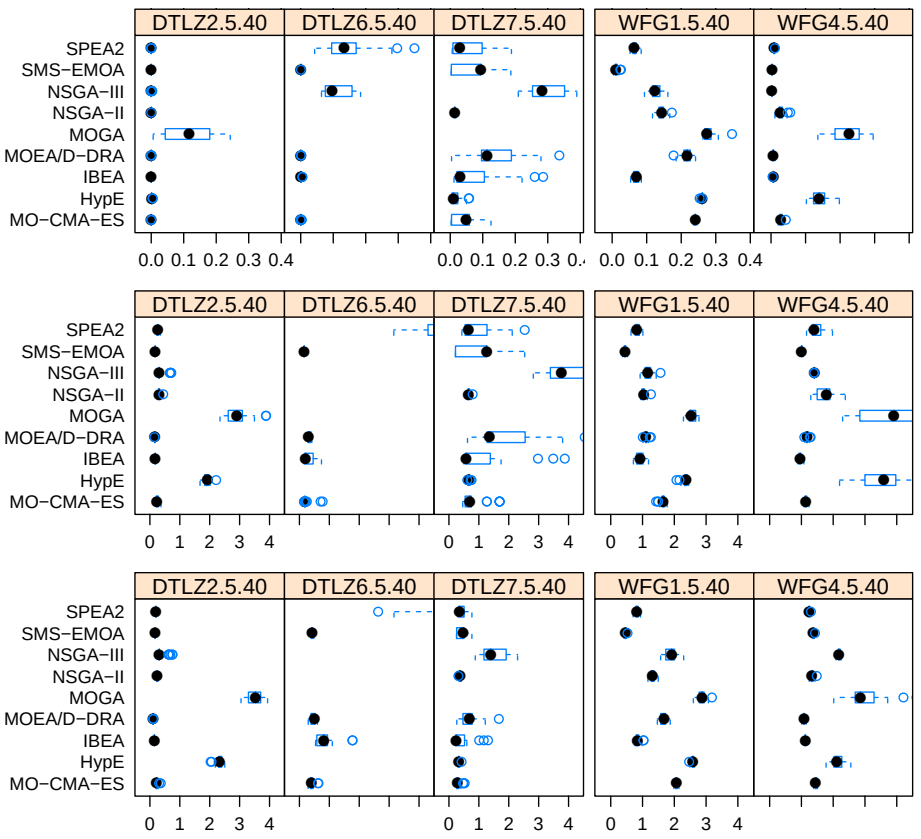

Figure 8: Performances of MOEAs given 40 000 FEs on selected five-objective problems with 40 variables. From top to bottom, I_H^{rpd}, I_ϵ and IGD.

Table 4: Sum of ranks (in parenthesis) depicting the performance of MOEAs on ten-objective problems.

2500 FEs									
I_H^{rpd}	**SMS** (0)	IBEA (827)	CMA (1919)	NSGA-II (2965)	NSGA-III (3543)	SPEA2 (4257)	HypE (5525)	MOEA/D (6218)	MOGA (7505)
I_ϵ	**MOEA/D** (0)	IBEA (370)	SMS (2092)	NSGA-II (2471)	CMA (2791)	NSGA-III (3315)	SPEA2 (3498)	HypE (4878)	MOGA (6297)
IGD	**IBEA** (0)	NSGA-III (847)	SPEA2 (1272)	CMA (1289)	SMS (1915)	NSGA-II (2228)	HypE (3315)	MOEA/D (4441)	MOGA (5562)
10000 FEs									
I_H^{rpd}	**IBEA** (0)	SMS (222)	CMA (1116)	NSGA-III (2326)	SPEA2 (2532)	NSGA-II (3241)	HypE (4846)	MOEA/D (5114)	MOGA (6919)
I_ϵ	**MOEA/D** (0)	IBEA (1258)	SMS (2794)	CMA (3250)	NSGA-III (3347)	NSGA-II (4045)	SPEA2 (4562)	HypE (5214)	MOGA (6922)
IGD	**NSGA-III** (0)	**IBEA** (36)	SPEA2 (646)	NSGA-II (1776)	SMS (1828)	CMA (1987)	HypE (2557)	MOEA/D (4424)	MOGA (5151)
40000 FEs									
I_H^{rpd}	**IBEA** (0)	SMS (897)	SPEA2 (1401)	CMA (1878)	NSGA-III (2416)	NSGA-II (2576)	HypE (4899)	MOEA/D (5077)	MOGA (7073)
I_ϵ	**MOEA/D** (0)	**IBEA** (220)	NSGA-III (1836)	SMS (2309)	NSGA-II (2986)	SPEA2 (3252)	CMA (3705)	HypE (4581)	MOGA (6518)
IGD	**IBEA** (0)	NSGA-III (928)	SPEA2 (942)	NSGA-II (2710)	HypE (2937)	CMA (3725)	SMS (4129)	MOEA/D (5513)	MOGA (6584)

On the non-concave WFG problems, many MOEAs show significant performance improvements, although it becomes clear that MOEAs are unable to converge to the actual fronts on these problems. Once again, results on WFG1 are fairly consistent across the different metrics. SMS ranks well on all these problems and across all metrics, but while it is the best-performing MOEA for WFG1–2, it is outperformed by MOEA/D on WFG3. In addition, IBEA also ranks well on WFG2. As for the concave WFG problems, we notice major discrepancies between performance metrics. For instance, SMS and NSGA-III rank very well according to the I_H^{rpd}, but on the distance-based metrics the performance of NSGA-III is not as competitive, and the same happens for SMS according to the *IGD*. Similarly, MOEA/D performs well according to the I_H^{rpd} and to the *IGD*, but is not as competitive according to the $I_{\epsilon+}^1$. Overall, the only MOEA that can be considered as well-performing according to all metrics is IBEA.

1.4 Ten-objective problems

The large increase in the number of objectives leads to important changes in the rank sum analysis we show in Table 4. First, as previously discussed, the disparity in rankings according to the different metrics becomes considerable. One could even say that selecting a best-performing MOEA becomes itself a multi-objective task, since only IBEA is able to rank well according to multiple metrics. Even so, other algorithms outrank it on specific metrics, corroborating that metrics play a critical role both for design and assessment on truly many-objective scenarios.

A very important observation that yet had not been reported in the literature is the low rank sums achieved by NSGA-II and SPEA2 in all ten-objective scenarios. Two precautions taken in this work are directly related to this good performance. First, both algorithms benefit directly from proper tuning, as the numerical parameters selected by irace differ considerably from the default adopted in the literature. Second, SPEA2 uses DE as underlying algorithm when given 10 000 and 40 000 FEs, reinforcing the need to consider different underlying EA algorithms when proposing a MOEA. Next, we detail our discussion for each budget considered.

1.4.1 2 500 FEs

Results depicted in the boxplots given in Fig. 9 show that the increase in the number of objectives makes several problems too difficult for MOEAs to solve when only a few FEs are allowed. More precisely, the only problem groups for which we see reasonable results are the moderate DTLZ problems (DTLZ2 and DTLZ4–5) and the non-concave WFG problems (WFG1–3). However, we remark that results for the I_ϵ on WFG3 in general are not as good as on WFG1–2.

On the selected problems depicted on Fig. 9, we see a contrast between metrics in opposite directions. While on DTLZ2 the performance of MOEAs look perfect, but the distance-based metrics show many differences between algorithms. In fact, the only MOEAs that consistently demonstrate good performance on the moderate DTLZ problems are SMS, IBEA, and MOEA/D. As for the hardest DTLZ problems, we see very poor performances from all MOEAs, with two major exceptions. First, I_H^{rpd} results for DTLZ6 when $n_{\text{var}} = 31$ look promising for SMS, IBEA, and MO-CMA-ES, but we remark that the distance-metrics disagree with this, except for SMS. Second, *IGD* results for SMS on DTLZ7 look reasonable, but the remaining metrics contradict these results.

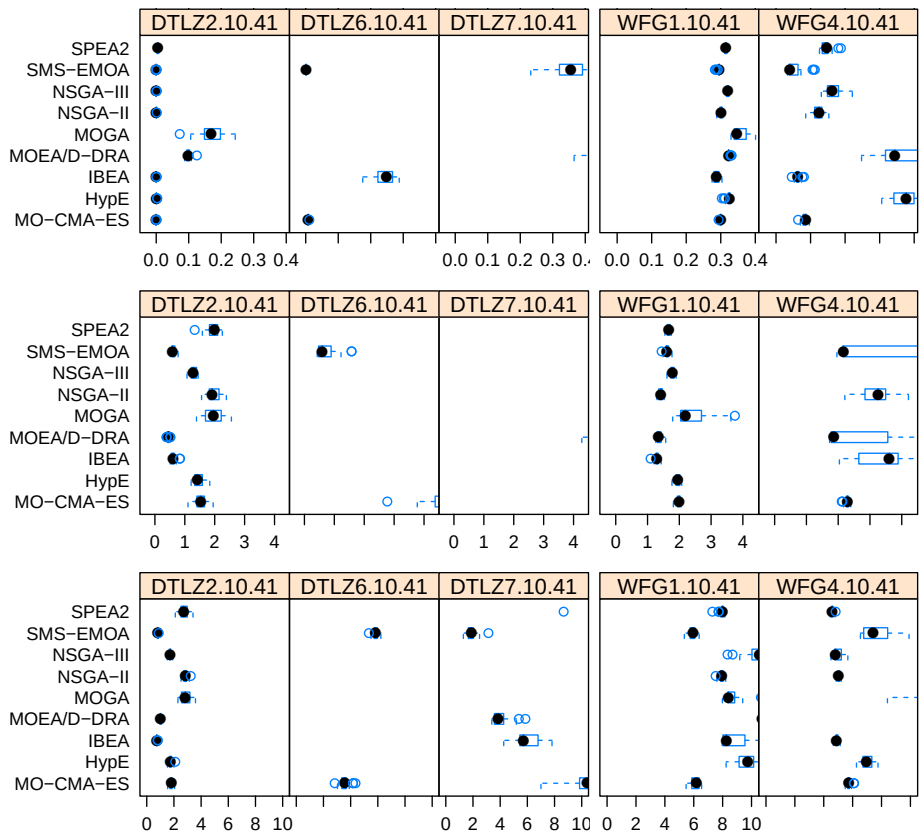

Figure 9: Performances of MOEAs given 2 500 FEs on selected ten-objective problems with 41 variables. From top to bottom, I_H^{rpd}, I_ϵ and *IGD*.

Concerning the non-concave WFG problems, we see three very different situations. First, on WFG1 the I_H^{rpd} and the *IGD* agree that results are far from reasonable, but the $I_{\epsilon+}^1$ indicates otherwise. By contrast, on WFG3 the indicator that suggests a good performance from MOEAs in general is the *IGD*. In this case however, the three metrics disagree completely, since the I_H^{rpd} would indicate a reasonable performance and the $I_{\epsilon+}^1$ results are very poor. Finally, the only of these problems where all metrics agree is WFG2. As a result, it is pretty difficult to select the best-ranked algorithm in all these problems. As for the concave WFG problems, we notice yet again that all metrics disagree. In particular, the distance-based metrics denote a very poor performance from all MOEAs, specially the *IGD*. In this context, it is more clear that some algorithms are unable to simultaneously satisfy multiple metrics than to indicate the best MOEA.

Overall, the major conclusion drawn from these experiments is that, when faced with computationally expensive problems that present a large number of objectives, MOEAs can only be expected to produce reasonable results for the ones that present particular features.

1.4.2 10 000 FEs

Results shown in Fig. 10 are very similar to the results discussed on the previous section, although improvements can be seen for most problems according to all metrics. On the moderate DTLZ problems, most MOEAs now get much closer to the actual fronts, the negative examples being the dominance-based algorithms. It is also important to remark that MOGA is surprisingly better on these functions than the remaining MOEAs from its paradigm according to the distance-based metrics. On DTLZ6, SMS and MO-CMA-ES now display good performance according to both I_H^{rpd} and $I_{\epsilon+}^1$, al-

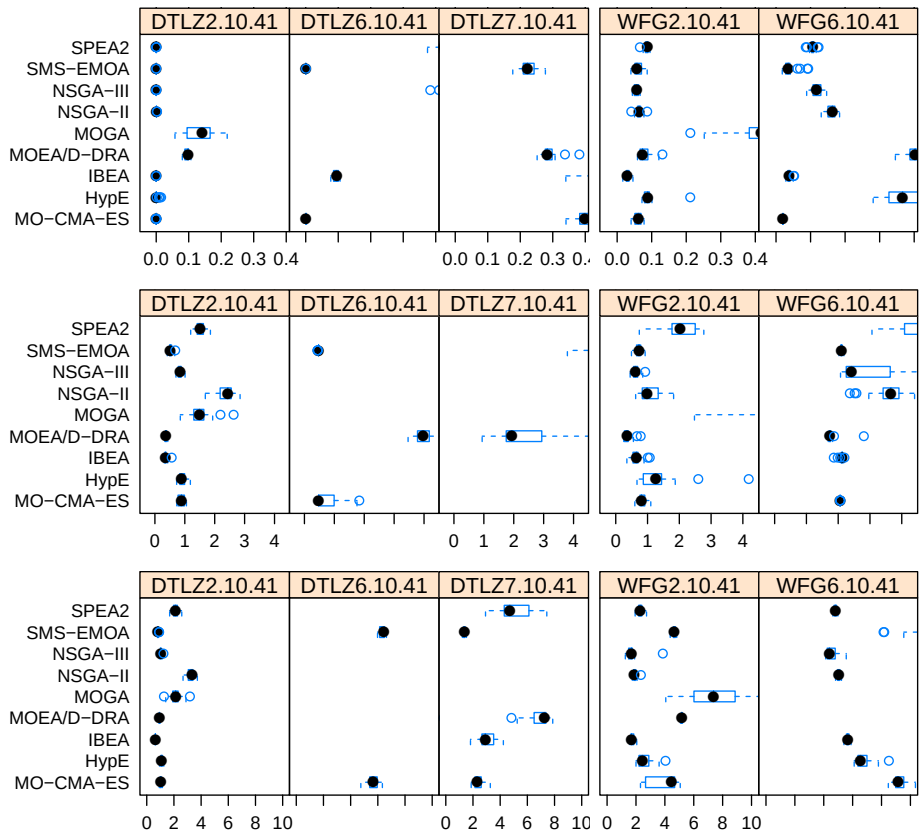

Figure 10: Performances of MOEAs given 10 000 FEs on selected ten-objective problems with 41 variables. From top to bottom, I_H^{rpd}, I_ϵ and *IGD*.

though mostly on smaller n_{var} values. On DTLZ7, $I_{\epsilon+}^1$ results are very poor for all MOEAs except for MOEA/D. By contrast, the only MOEA one could recommend based on the I_H^{rpd} and the *IGD* is SMS.

Concerning the non-concave WFG problems, results are again unique. For WFG1, the only metric that points to a reasonable performance from MOEAs in general is the $I_{\epsilon+}^1$. As for WFG2, metrics again consistently disagree, with IBEA, MOEA/D, and NSGA-III being the algorithm of choice for each metric, respectively. Nonetheless, we remark that results from half of the MOEAs in this problem are quite good, even if they are not able to fully approximate the actual front. Finally, on WFG3 results are still far from good, but can be considered much better than the ones for WFG1. Surprisingly, the MOEA of choice for this problem is NSGA-II.

As for the non-concave problems, once again it is not possible to recommend a single MOEA due to the disparity between different metric rankings. Given the I_H^{rpd}, the $I_{\epsilon+}^1$, and the *IGD*, the algorithms of choice would be SMS, MOEA/D, and NSGA-III, respectively.

1.4.3 40 000 FEs

The increase in the number of FEs given to MOEAs this time results in several significant performance improvements only for the I_H^{rpd}, as seen in Fig. 11. In particular, we remark that all MOEAs benefit from the increase FE budget on most problems. The only exception concerns the WFG1, where IBEA, MOEA/D, and MO-CMA-ES are unable to improve over their previous results. Concerning the distance-based metrics, improvements can be seen only on the moderate DTLZ functions, although the *IGD*

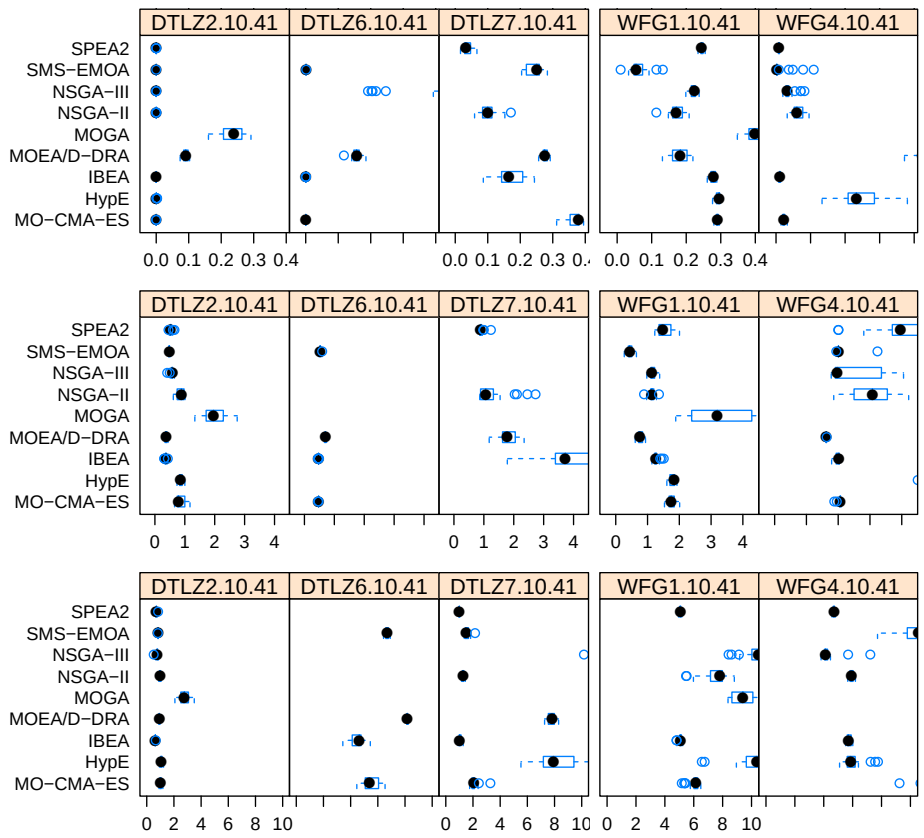

Figure 11: Performances of MOEAs given 40 000 FEs on selected ten-objective problems with 41 variables. From top to bottom, I_H^{rpd}, I_ϵ and *IGD*.

also shows minor improvements for the hardest functions. We remark, though, that we observe in a few occasions that an improvement according to a given metric might result in a worsening on another. Such is the case with WFG1, for instance, where the algorithms that improve the most on the I_H^{rpd} worsen on the *IGD*.

Overall, results indicate that even given a large number of FEs no single MOEA is able to perform consistently well on all problems. In addition, it happens very often that the best MOEA for a given problem is only considered the best for that particular problem, or that MOEAs are only the best-performing algorithms according to a given metrics. Altogether, these results suggest that the task of a general-purpose MOEA for unconstrained continuous is probably unfeasible when dealing with truly many-objective optimization.