OCD: Online Convergence Detection for Evolutionary Multi-Objective Algorithms Based on Statistical Testing

Tobias Wagner[1], Heike Trautmann[2], and Boris Naujoks[3]

[1] Institute of Machining Technology (ISF), TU Dortmund University, Germany
wagner@isf.de
[2] Faculty of Statistics, TU Dortmund University, Germany
trautmann@statistik.tu-dortmund.de
[3] Chair of Algorithm Engineering, TU Dortmund University, Germany
boris.naujoks@tu-dortmund.de

Abstract. Over the last decades, evolutionary algorithms (EA) have proven their applicability to hard and complex industrial optimization problems in many cases. However, especially in cases with high computational demands for fitness evaluations (FE), the number of required FE is often seen as a drawback of these techniques. This is partly due to lacking robust and reliable methods to determine convergence, which would stop the algorithm before useless evaluations are carried out. To overcome this drawback, we define a method for online convergence detection (OCD) based on statistical tests, which invokes a number of performance indicators and which can be applied on a stand-alone basis (no predefined Pareto fronts, ideal and reference points). Our experiments show the general applicability of OCD by analyzing its performance for different algorithmic setups and on different classes of test functions. Furthermore, we show that the number of FE can be reduced considerably – compared to common suggestions from literature – without significantly deteriorating approximation accuracy.

1 Introduction

In real-world industrial problems and engineering applications, improvements, e.g., in simulation techniques, machines, tools, and materials, constantly offer increasing productivity. However, in order to completely exploit these potentials, an appropriate setup of the inherent parameters is necessary. Due to the numerous requirements of modern processes, these problems are mainly multi-objective, which supports the application of evolutionary multi-objective algorithms (EMOA). Nevertheless, their applicability is still put into question, even though EMOA have already been successfully applied to these kinds of problems.

A possible reason, for instance when compared to mathematical programming methods, may be the lack of convergence criteria for EMOA. More specific, the performance of an a-posteriori multi-objective optimization algorithm can be expressed in simple terms by two objectives:

M. Ehrgott et al. (Eds.): EMO 2009, LNCS 5467, pp. 198–215, 2009.

1. maximize the quality of the Pareto-front approximation and
2. minimize the number of function evaluations or computation time, respectively.

In the last decade, many EMOAs have been introduced to achieve one or both of the above objectives. For instance, the use of performance indicators [1,2,3], which evaluate the quality of the current Pareto-front approximation, has turned out to be successful in achieving the first objective [4]. The second objective has recently been approached by integrating modeling methods into the EMOA framework [5,6,7]. However, in the evaluation of all these methods, the number of allowed function evaluations (FE) is fixed at a predefined level, which is high (30k-500k FE [8,9]) when the main objective is a good approximation and low for model-assisted approaches (130-250 FE [6,7]). In order to perform the optimization in an efficient manner, the EMOA should be stopped when

1. no improvement can be gained by further iterations or
2. the approximation quality has reached the desired level.

Right now, these stopping criteria are only applied for single-objective approaches. Nevertheless, the detection of convergence is an equally important issue for EMOA since further evaluations are a waste of computational resources and may lead to a loss of diversity by means of genetic drift [10]. Multi-objective performance indicators allow the reduction of a multi-objective optimization (MOO) problem to a single-objective problem on sets [3]. Thereby, the above criteria can be transferred to MOO. Furthermore, multiple indicators can be used to reliably detect different kinds of improvement in the set.

In this paper, an approach for online convergence detection (OCD) is introduced. Due to the stochastic nature of evolutionary algorithms, OCD is based on systematic statistical testing. The number of parameters is low, it can be combined with any set-based EMOA, and the selection of the considered preference indicators is up to the user. Thus, OCD is an intuitive, yet flexible tool to guarantee an effective use of EMOA, which may promote the industrial application of these methods.

In section 2, the state of the art in multi-objective convergence detection is summarized. Afterwards, OCD is detailed, and the algorithmic steps are presented (section 3). The applicability of OCD is demonstrated by comprehensive experiments, which are described and analyzed in section 4. Finally, conclusions are drawn and the results are summarized in section 5.

2 State of the Art

For the application of EMOA on new industrial problems, where no sufficient a-priori knowledge exists, it is generally hard to find a suitable termination criterion. Therefore, the most frequently used limit is the maximum number of generations or FE. Hybrid EMOA using quadratic programming methods have been developed to guarantee (local) optimality of solutions [11,12]. These approaches are formally converged as soon as Karush-Kuhn-Tucker (KKT) points

for a given set of aggregation or reference-point-based distance functions have been identified, but can not guarantee the quality of the set of solutions, e.g., in terms of diversity and spread. This is accomplished by recent approaches, which compute the gradient of the hypervolume for a set of solutions [13]. Note that all these approaches require sufficient accuracy in the approximation of the Hessian.

Deb and Jain [14] investigate so-called running performance metrics for convergence and diversity of solutions to be monitored in the course of the algorithm. Thereby, the algorithm may be stopped when convergence is observed. However, therein the authors focus on performance evaluation and algorithm comparison. An automated procedure for detecting convergence has not been proposed. For this purpose, Rudenko and Schoenauer [15] survey possible online termination criteria for elitist EMOA, such as the disappearance of all dominated individuals or the deterioration of the number of newly produced non-dominated individuals. Finally, they suggest a technique for determining stagnation based on stability of the maximum crowding distance, which requires the determination of a threshold, which depends on the scale of the objectives as well as the population size. Furthermore, its application is only tested with NSGA2, which uses the crowding distance as selection criterion [16]. It is an open question whether a stability of the maximum crowding distance can be observed in EMOA, which do not directly use this measure in the selection process.

The basic idea of using dominance-related metrics to compare sets [17] has recently been used to reduce the multi-objective to a single-objective problem on sets [3]. This allows to use convergence criteria from single-objective theory. Furthermore, a method for offline detection of the expected generation, in which the EMOA converges, has been introduced [18]. This method is based on statistical testing of the similarity in the distribution of performance measures for consecutive generations relying on multiple parallel runs of the EMOA. In this paper, the main ideas of both contributions are transferred to online convergence detection.

3 Online Convergence Detection

In the progression of OCD, two different analyses are carried out. It is sequentially tested whether the variance of the performance indicator values decreases below a predefined limit ($VarLimit$) or whether no significant trend of the performance indicators can be detected over the last generations. The EMOA terminates if at least one of these conditions is met.

All algorithmic steps of the proposed OCD approach and the required subroutines are given in Algorithms 1, 2, and 3. These steps are described in depth to ensure a straightforward implementation of OCD. The required input parameters for Algorithm 1 can be set easily, even by inexperienced users. The variance limit $VarLimit$ corresponds to the desired approximation accuracy in single-objective optimization, but does not require knowledge about the actual minima of the objectives. The algorithm stops when the standard deviation of the indicator values over the given time window of $nPreGen$ generations is significantly below $\sqrt{VarLimit}$. Thus, the user can exactly determine how many

Algorithm 1. OCD: Algorithm for Online Convergence Detection

Require: $VarLimit$ /* *maximum variance limit* */
$nPreGen$ /* *number of preceding generations for comparisons* */
α /* *significance level of the tests* */
$MaxGen$ /* *maximum generation number* */
$(PI_1, \ldots, PI_n)$ /* *vector of performance indicators, e.g., (HV, ε, R2)* */

1. $i = 0$ /* *initialize generation number* */
2. **for all** $i \in \{1, \ldots, nPreGen\}, j \in \{1, \ldots, n\}$ **do**
3. $pChi2(j, i) = 1$ /* *initialize p-values of the χ^2-variance Test* */
4. $pReg(i) = 0$ /* *initialize p-values of the t-Test on regression coefficient* */
5. **end for**
6. $\boldsymbol{lb} = []$ /* *initialize lower bound vector* */
7. $\boldsymbol{ub} = []$ /* *initialize upper bound vector* */
8. **repeat**
9. $i = i + 1$
10. Compute d-objective Pareto front PF_i of i-th EMOA generation
11. $\boldsymbol{lb} = min(\boldsymbol{lb} \cup PF_i)$ /* *update lower bound vector* */
12. $\boldsymbol{ub} = max(\boldsymbol{ub} \cup PF_i)$ /* *update upper bound vector* */
13. **if** $(i > nPreGen)$ **then**
14. $PF_i = 1 + (PF_i - \boldsymbol{lb})/(\boldsymbol{ub} - \boldsymbol{lb})$ /* *normalize PF_i to $[1, 2]^d$* */
15. **for all** $k \in \{i - nPreGen, \ldots, i - 1\}$ **do**
16. Compute Pareto front PF_k of k-th EMOA generation
17. $PF_k = 1 + (PF_k - \boldsymbol{lb})/(\boldsymbol{ub} - \boldsymbol{lb})$ /* *normalize PF_k to $[1, 2]^d$* */
18. **end for**
19. **for all** $j \in \{1, \ldots, n\}$ **do**
20. $\boldsymbol{PI}_{j,i} = (PI_j(PF_{i-nPreGen}, PF_i, \mathbf{1}, \mathbf{2.1}), \ldots, (PI_j(PF_{i-1}, PF_i, \mathbf{1}, \mathbf{2.1})))$
/* *compute PI_j for $PF_{i-nPreGen}, \ldots, PF_{i-1}$ using PF_i as reference set, $\mathbf{1}$ as ideal, and $\mathbf{2.1}$ as reference point* */
21. $pChi2(j, i) =$ **call** $Chi2(\boldsymbol{PI}_{j,i}, VarLimit)$ /* *p-value of χ^2 test* */
22. **end for**
23. $pReg(i) =$ **call** $Reg(\boldsymbol{PI}_{1,i}, \ldots, \boldsymbol{PI}_{n,i})$
/* *p-value of the t-Test on the generation's effect on the $\boldsymbol{PI}_{j,i}$* */
24. **end if**
25. **until** $\forall j \in \{1, \ldots, n\} : (pChi2(j, i) \leq \alpha/n) \wedge (pChi2(j, i-1) \leq \alpha/n)$
$\vee \quad (pReg(i) > \alpha) \wedge (pReg(i-1) > \alpha)$
$\vee \quad i = MaxGen$
26. Terminate EMOA
27. **return** $\{MaxGen, Chi2, Reg\}$ /* *criterion which terminates the EMOA* */
i /* *generation in which the criterion holds* */

generations the EMOA is maximally allowed to compute with average changes in the indicator values significantly below the specified limit. The user also has to specify a significance level α for each statistical test procedure. Established levels for α, such as 0.05 (standard) and 0.01 (conservative), exist. The maximum generation number $MaxGen$ ensures that the resources required by the algorithm cope with the restrictions of the individual application, especially in the case where no convergence of the EMOA can be detected. However, the maximum number of function evaluations has to be specified for most known EMOA

as well. The number and types of desired performance indicators (PI) have to be selected in order to evaluate the solution quality at each generation with respect to the requirements of the user, which allows him to express his own preferences on the final Pareto-front approximation [3]. Users, who are not familiar with multi-objective performance assessment, can resort to the standard set of PI as defined by Knowles et al. [19], which comprises the hypervolume, the additive ε-, and the R2 indicator. Only these indicators meet the requirement of strict compliance with the Pareto dominance relation.

After the first $nPreGen$ generations, convergence is checked after each generation i. The n indicator values of the vector $\boldsymbol{PI}_{j,i}$ $(j = 1, \ldots, n)$ are computed for the objective sets of generations $i - nPreGen, \ldots, i - 1$ using the Pareto-front approximation of generation i as reference set. Thus, no a-priori knowledge about the true Pareto front is required, making the method applicable to practical problems. If a specific indicator PI_j does not use a reference set and evaluates each set separately (e.g., the hypervolume indicator), the difference between the indicator value of the preceding and the current set is calculated and stored in $\boldsymbol{PI}_{j,i}$.

The sets are normalized to the interval $[1, 2]^d = [1, 2] \times \ldots \times [1, 2] \subset \mathbb{R}^d$ as it is also implemented in PISA [20], where d is the number of objective dimensions. This is done in order to avoid problems within the indicator calculation based on objectives which are negative, equal to zero, or extremely large [19]. Since the actual bounds of the non-normalized objectives are not a-priori known, they are updated at each generation. The Pareto-front approximations of the $nPreGen$ preceding generations are also normalized based on the current objective-bound approximations. Due to the normalization, $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^d$ and $\mathbf{2.1} = (2.1, \ldots, 2.1) \in \mathbb{R}^d$ can be used as ideal point and (anti-optimal) reference point for the PI calculation, respectively.

The resulting $nPreGen$ vectors of n indicator values at each generation are then – separately for each indicator – checked against the alternative hypothesis that the variance of these values is lower than the predefined threshold $VarLimit$ using the χ^2-variance test [21] (cf. Algorithm 2). This parametric test is used being aware of its sensitivity to the normality assumption of the underlying sample as no nonparametric test for this problem exists. Due to the multiple testing, a Bonferroni correction on α is performed [22] resulting in an individual significance level of α/n for each test. The α-correction ensures that at each generation the global desired significance level is met. However, a correction with respect to the sequential testing over all generations is impossible concerning a reasonable applicability of OCD.

Additionally, a regression analysis is performed in order to check the significance of the descending linear trend (cf. Algorithm 3). Unfortunately, a test for $H_0 : \beta \neq 0$ vs. $H_1 : \beta = 0$ cannot be constructed. Thus, the test has to be performed with interchanged hypotheses, and the generation, in which the null hypothesis cannot be rejected anymore, has to be determined. Additionally, the decreasing linear trend has been checked via the negative sign of the estimator $\hat{\beta}$.

Algorithm 2. *Chi*2: One-Sided χ^2-variance test for

$$H_0: \quad \mathrm{var}(\boldsymbol{PI}) \geq VarLimit \quad \text{vs.} \quad H_1: \quad \mathrm{var}(\boldsymbol{PI}) < VarLimit$$

Require: $\boldsymbol{PI}$ /* vector of performance indicator values */
$VarLimit$ /* variance limit */
1. $N = length(\boldsymbol{PI}) - 1$ /* determine degrees of freedom */
2. $Chi = [\mathrm{var}(\boldsymbol{PI}) * N]/VarLimit$ /* compute test statistic */
3. $p = \chi^2(Chi, N)$ /* look up χ^2 distribution function with N degrees of freedom */
4. **return** p

Strictly speaking, the α-error for the desired decision cannot be controlled by α, but equals $1 - power(\text{t-test})$, where the *power* of a statistical test is the probability that the test will reject a false null hypothesis. As a result, an overall significance level at generation i cannot be maintained since the χ^2-variance test initiates the EMOA termination in the case of H_0 being rejected whereas the t-test initiates it in the opposite case. Thus, no combination of the α-levels can be performed relating to multiple test theory [22] although both tests are simultaneously performed on the same data. However, the main focus when setting up α is not on correctly controlling the α-error, but on finding reasonable critical values for the test statistics in order to make OCD applicable and successful within industrial applications.

Algorithm 3. *Reg*: Two-sided t-test on the significance of the linear trend

$$H_0: \quad \beta = 0 \quad \text{vs.} \quad H_1: \quad \beta \neq 0$$

Require: $\boldsymbol{PI}_j, \quad j = (1, \ldots, n)$ /* vectors of performance indicator values */
1. $N = n \cdot length(\boldsymbol{PI}^*) - 1$ /* determine degrees of freedom */
2. **for all** $j \in \{1, \ldots, n\}$ **do**
3. $\quad \boldsymbol{PI}^*{}_j = (\boldsymbol{PI}_j - \overline{\boldsymbol{PI}_j})/\sigma_{\boldsymbol{PI}_j}$ /* standardize */
4. **end for**
5. $\boldsymbol{PI}^* := \mathrm{concatenate}(\boldsymbol{PI}^*{}_1, \ldots, \boldsymbol{PI}^*{}_n)$ /* row vector of all $\boldsymbol{PI}_j$ */
6. $X = \underbrace{(1, \ldots, length(\boldsymbol{PI}^*), \ldots, 1, \ldots, length(\boldsymbol{PI}^*))}_{n \text{ times}}$
/* row vector of generations corresponding to $\boldsymbol{PI}^*$ */
7. $\hat{\beta} = (X * X^T)^{-1} * X * (\boldsymbol{PI}^*)^T$ /* linear regression without intercept */
8. $\varepsilon = \boldsymbol{PI}^* - X * \hat{\beta}$ /* compute residuals */
9. $s^2 = (\varepsilon * \varepsilon^T)/N$ /* mean squared error of regression */
10. $t = \frac{\hat{\beta}}{\sqrt{s^2 (X * X^T)^{-1}}}$ /* compute test statistic */
11. $p = 2 \cdot min(\mathrm{t}_N(t), 1 - \mathrm{t}_N(t))$
/* look up p-value from t distribution with N degrees of freedom */
12. **return** p

For performing the t-test, all indicator values $\boldsymbol{PI}_j$ are standardized, i.e., linearly transformed to mean zero and standard deviation one. The standardization

of $\boldsymbol{PI}_j$ provides two benefits: the regression can be performed for all indicators at once and no intercept (constant term) is required. The least squares estimator $\hat{\beta}$ of the actual slope β is determined in line 7 [23]. Afterwards, the fit is calculated via the mean squared error of the linear model, and a standard error of the estimator is computed [23]. Based on these measures, the t-statistic, i.e., the standardized regression coefficient, and the p-value can be computed using a standard statistical library.

The algorithm stops if either the variance test or the regression analysis indicates the convergence of the EMOA for generations i and $(i-1)$. OCD returns the stopping generation i and the method that initiated the EMOA termination. Thereby, the user is informed about the final state of the algorithm. In the case of termination based on the maximum number of generations, the user knows that the EMOA has not yet converged and further generations may further improve the Pareto-front approximation.

Additional Runtime for OCD

The update, normalization, and standardization of the objective sets within each iteration can be performed in $O(N)$, where N denotes the population size. The calculation of the Pareto front requires $O(N \log^{d-1} N)$ [24], but is already part of most known EMOA. Thus, the calculation of the indicator values is the crucial part of OCD. Especially when the hypervolume is used, the runtime is in $O(N^{d/2+1})$ for $d > 3$ [25]. For hypervolume-based algorithms, such as SMSEMOA [2], this is not critical since the selection procedure is in the same complexity as OCD. Also for expensive real-world problems, the time, which can be saved by an appropriate termination, is considerably higher than the additional runtime. Nevertheless, the approach can be efficiently used for time-critical optimization as well by using performance measures in $O(Nd)$, such as the R2 indicator.

4 Experiments

The experiments are conducted to analyze the proposed OCD applied to modern EMOA. At present, online convergence detection can only be performed by a human decision maker, who inspects the running metrics, i.e., the PI, and terminates the algorithm when convergence is observed. For a successfully automatized application, the time when OCD indicates convergence has to be in agreement with the intuitive understanding of the decision maker. Thus, the first experiments focus on the correspondence of OCD and a human decision maker. In order to analyze the applicability of the statistical tests separated from the whole OCD framework, OCD is additionally computed using pre-calculated Pareto front discretizations as well as the known ideal and anti-ideal points. Apart from the OCD version in Algorithm 1, we will refer to the latter as OCD with full information. Finally, the results received by standard OCD are compared to the common termination criterion from EMOA literature, i.e., a fixed

number of FE. Here, we focus on the reduction of the number of evaluations as well as the loss of quality by stopping the evolution earlier.

Research Question. The main question of the analysis is whether or not the proposed OCD algorithm helps to reduce FE without resulting in an uncontrollable loss of quality. Therefore, we evaluate the results received regarding both approximation quality and the required number of FE and compare them to the ones we receive after applying the number of FE, which are originally proposed in standard EMO literature. Moreover, we are interested in the criterion which first indicates convergence and how this is motivated by the $\boldsymbol{PI}_{i,j}$ characteristics over time. In order to inspect the behavior of OCD more closely, it is also analyzed whether OCD, with the reference set and the ideal and anti-ideal point approximated on the fly, performs similar to the case of full information. Last but not least, we want to demonstrate that the time, when OCD indicates convergence, matches with an intuitive observation of the running metrics.

Pre-experimental planning. NSGA2 [16] and SMSEMOA [2] are considered since NSGA2 is the industrially most popular EMOA and recent studies motivate the use of the hypervolume contribution during selection [4]. The test functions are chosen to represent different kinds of problem characteristics, such as dimension in decision and in objective space, the number of local optima, and the shape of the Pareto front. The population sizes used on the problems vary in order to allow for different problem characteristics and evaluate OCD for a wider variety of algorithmic setups.

Initial preparative analyses of OCD indicate that the time window $nPreGen$ should span at least seven, but better ten, generations to permit an adequate calculation of the p-values in the tests. In this context, it has to be considered that the tests will not indicate convergence until the $\boldsymbol{PI}_{j,i}$ stagnate over a large span of this time window. Thus, when it is reviewed whether OCD's indication matches with the generation determined by a human decision maker, the delay of $nPreGen$ generations has to be accepted within the assessment.

Task. Check if OCD provides a robust and reliable termination of EMOA on several test cases. Compare the results of OCD with an intuitive understanding of termination and with the results provided in standard EMO literature. Furthermore, systematical deviations between the proposed approach and the one with full information are to be identified, which may occur due to a inaccurate approximation of the true Pareto front.

Setup. NSGA2 and SMSEMOA are analyzed on the four bi-objective test functions Fonseca [27], ZDT1, ZDT2, and ZDT4 [28] as well as on the three-objective DTLZ2 [29] test function. Different population sizes $\mu \in \{60$ (Fonseca), 100 (ZDT1, ZDT2, DTLZ2), 200 (ZDT4)$\}$ and selection strategies – $(\mu + \mu)$ in the NSGA2 and $(\mu+1)$ in the SMSEMOA – are incorporated, where, for the sake of comparability, a generation of SMSEMOA equals a sequence of μ FE. For each combination of EMOA and test function, ten independent runs are performed.

The variance bound for the χ^2-variance test is set to $VarLimit = 0.001^2$, the significance level for both tests is set to $\alpha = 0.05$, and the time window is of size

Table 1. Parameter settings within the experiments

test problem	$MaxGen$	$MaxGen2$
Fonseca	66	66
ZDT1, ZDT2	120	200
ZDT4	200	100
DTLZ2	300	300

algorithm	implem.	p_c	p_m	η_c	η_m	p_{swap}
NSGA2	R [26][a]	0.7	0.2	20	20	0
SMSEMOA	PISA [20]	0.9	1/length($\boldsymbol{x}$)	15	20	0.5

[a] NSGA2 is taken from the package *mco* (http://cran.r-project.org/web/packages/mco/index.html).

$nPreGen = 10$. The different numbers of FE allowed within our experiments ($MaxGen$) and within the standard literature (MaxGen2) [8] as well as the parameters used in the simulated binary crossover and polynomial mutation [30] are displayed in Tab. 1. For measuring the performance of the algorithms, the following PI have been invoked: hypervolume (HV) [31], additive ε (Eps) [17], and R2 [32]. Recall that OCD as well as OCD with full information terminate if and only if one of the tests (χ^2-variance or t-test) simultaneously indicates convergence with respect to all three metrics. The reference fronts used within OCD with full information have been calculated via equidistant sampling of the known Pareto fronts.

Experimentation/Visualization. Several ways of visualization are used to demonstrate our findings. In the first plots, the PI behavior is inspected over the generations of the EMOA on the ZDT4 (cf. Fig. 1) and the DTLZ2 test function (cf. Fig. 2), where the median run with respect to the difference between the full information-based performance metrics and OCD is plotted semi-logarithmically. The black and light-gray solid lines indicate the generation, in which either the χ^2-variance or the regression criterion detect convergence in case of the reference set and objective bounds being approximated online. The black and light-gray dashed lines indicate the generation, in which convergence is detected for the given combination of EMOA and test problem within the full information approach.

The differences in performance are visualized using boxplots. The subsequent figures present the differences between the $\boldsymbol{PI}_{j,i}$ after the number of FE recommended in literature ($i = MaxGen2$) and after OCD indicated convergence. One box is shown for each PI_j and each considered test case, in Fig. 3 for the NSGA2 and in Fig. 4 for the SMSEMOA. Due to different scales, the displayed area had to be changed for some of the test cases, i.e., DTLZ2 for NSGA2 and ZDT1 as well as ZDT2 for the SMSEMOA. For the combinations of EMOA and test function, in which the variance criterion initiated termination for most of the runs, the interval $[-\sqrt{VarLimit}], \sqrt{VarLimit}]$ is highlighted in order to assist inspecting the effect of $VarLimit$ on the final approximation quality. Fig. 5 splits the runs for all test problems into two categories: runs being terminated by the regression criterion and by the χ^2-variance test. This analysis is done separately for NSGA2 and SMSEMOA in order to show the two different types of EMOA behavior and how OCD copes with these challenges.

Statistic details of the boxplots can be found in Tab. 2. Here, the median differences are listed with respect to the corresponding algorithm/test case combination. Note that all median differences are given multiplied by 10^{-3}. Besides

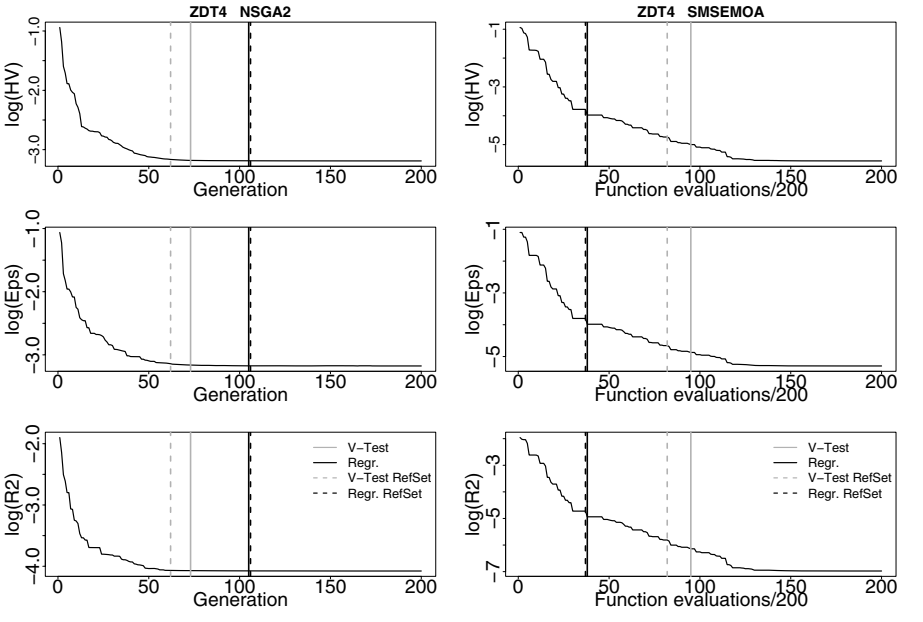

Fig. 1. The run of the metrics with respect to the reference set for NSGA2 and SMSEMOA on ZDT4. Exemplary, the run is chosen which obtains the median difference between the approaches with and without full information. The vertical lines indicate the generations, in which the different tests and variants of OCD would stop the algorithm.

the results regarding the received quality, the additional rows within Tab. 2 indicate the number of generations OCD terminated the algorithm earlier in contrast to the generation number suggested in the literature ($MaxGen2$) [8]. Furthermore, the number of saved function evaluations and their percentage of MaxGen2 are calculated to emphasize what is saved by using OCD with only the given median loss in quality.

In the line plots of Fig. 6 and Fig. 7, the values of each run with and without full information are compared. By these means, systematic deviations can easily be observed. Since OCD terminates the EMOA when the first of the tests indicates convergence, it is also labeled which of the tests initiates the termination of each run using different symbols. Fig. 6 shows the results for NSGA2 on each test case whereas Fig. 7 provides these for SMSEMOA.

Observations. OCD efficiently copes with two different types of convergence. In case the variance test terminates the EMOA (cf. Fig. 5, subfigures 1 and 3), the standard deviation of all $\boldsymbol{PI_{i,j}}$ is significantly below $\sqrt{VarLimit} = 0.001$. Fig. 5 shows that the $\boldsymbol{PI_{j,i}}$ differences between OCD Stop and $MaxGen2$ are approximately in the range of $[-0.001, 0.001]$ for the EMOA runs, which have been terminated by the χ^2-variance-test. Furthermore, big differences to the runs, which are terminated by the regression criterion (cf. Fig. 5, subfigures 2

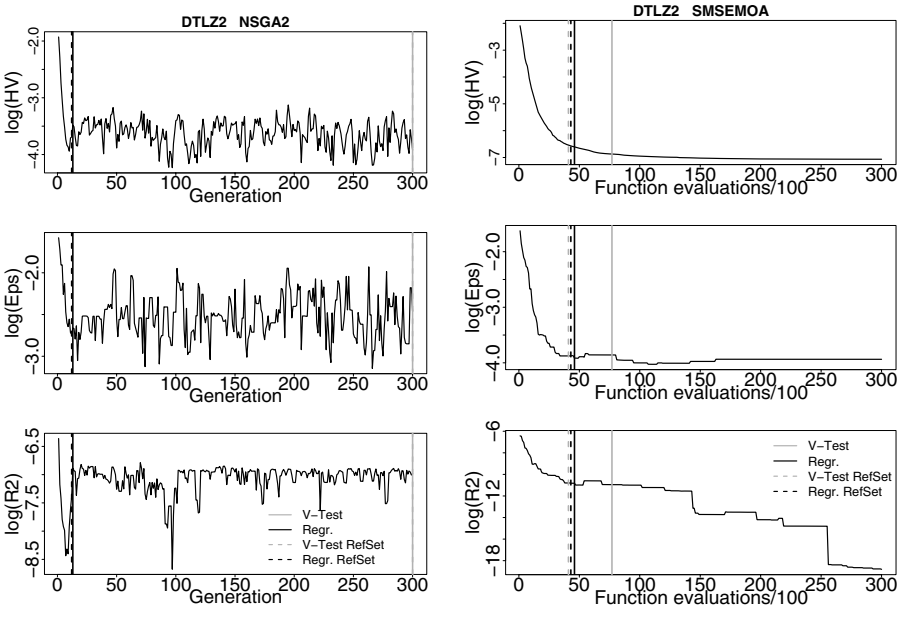

Fig. 2. The run of the metric with respect to the reference set for NSGA2 and SMSEMOA on DTLZ2. Exemplary, the run is chosen which obtains the median difference between the approaches with and without full information. The vertical lines indicate the generations, in which the different tests and variants of OCD would stop the algorithm.

and 4), can be observed. In these cases the differences between the approximation quality of OCD Stop and $MaxGen2$ are much higher, strictly positive for the SMSEMOA and balanced between positive and negative values for NSGA2.

The basic results from above can also be recognized in the boxplots for NSGA2 (cf. Fig. 3) and SMSEMOA (cf. Fig. 4). However, systematic differences between the NSGA2 and the SMSEMOA results can be detected on ZDT4 and DTLZ2. For NSGA2 on ZDT4, the variance criterion indicates convergence much earlier than the regression criterion. This is different from the findings for SMSEMOA, where the regression criterion terminates the algorithm earlier. The progressions of $\boldsymbol{PI}_{j,i}$ on DTLZ2 are strongly distorted for NSGA2 with alternating phases of convergence and divergence. The ones of SMSEMOA are much smoother. In both cases, the regression criterion is able to identify convergence very early in the run, but due to the rough structure, the variance test is not able to do so for NSGA2, while for SMSEMOA the variance criterion terminates the optimization about 25 to 30 generations later than the regression criterion.

The differences in generations between the ones proposed by OCD and $MaxGen2$ range from rather small (18 for NSGA2 on ZDT4) to very large (287 for NSGA2 on DTLZ2). In the latter case, only less than 5% of the evaluations are needed to find better solutions compared to the ones found after the complete optimization run with the termination criterion proposed in the literature.

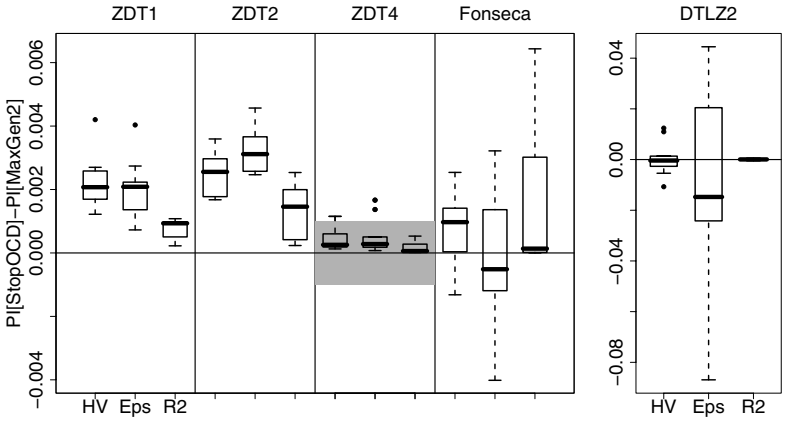

Fig. 3. Boxplots of PI differences at OCD StopGen and $MaxGen2$ for NSGA2. The interval $[-\sqrt{VarLimit}], \sqrt{VarLimit}]$ is highlighted in gray where appropriate.

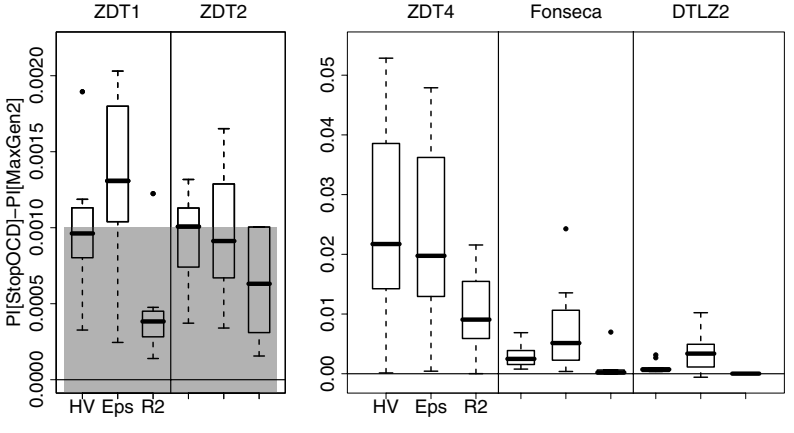

Fig. 4. Boxplots of PI differences at OCD StopGen and $MaxGen2$ for SMSEMOA. The interval $[-\sqrt{VarLimit}], \sqrt{VarLimit}]$ is highlighted in gray where appropriate.

In most cases, slightly more than 50% of the generations can be saved. This results in over 10,000 unnecessary evaluations for the high-dimensional problems. Even in the worst case, more than 2,900 evaluations can be saved.

The coincidence of both tested OCD variants are indicated in the line plots in Fig. 6 for NSGA2 and Fig. 7 for SMSEMOA. The differences between OCD and its full-information variant are strongly depending on the EMOA in use. For SMSEMOA the results with full information and approximated reference sets are well-correlated and no general trend can be observed. The median differences between the indications of convergence in both situations are within one to five generations (cf. Fig. 7). This is different to NSGA2, which shows a trend to overestimate the stop generation for the high-dimensional problems. Furthermore, some outliers

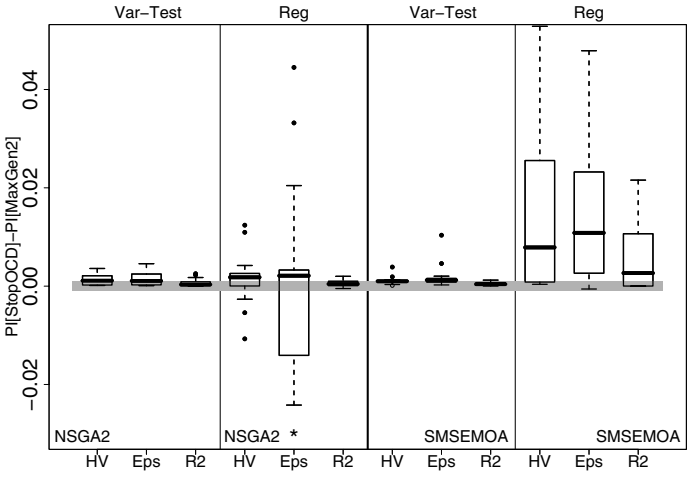

Fig. 5. Separated boxplots of PI differences between OCD StopGen and $MaxGen2$ for the runs of the SMSEMOA which are terminated by the χ^2-variance test and the regression analysis, respectively. *Two extreme outliers not shown.

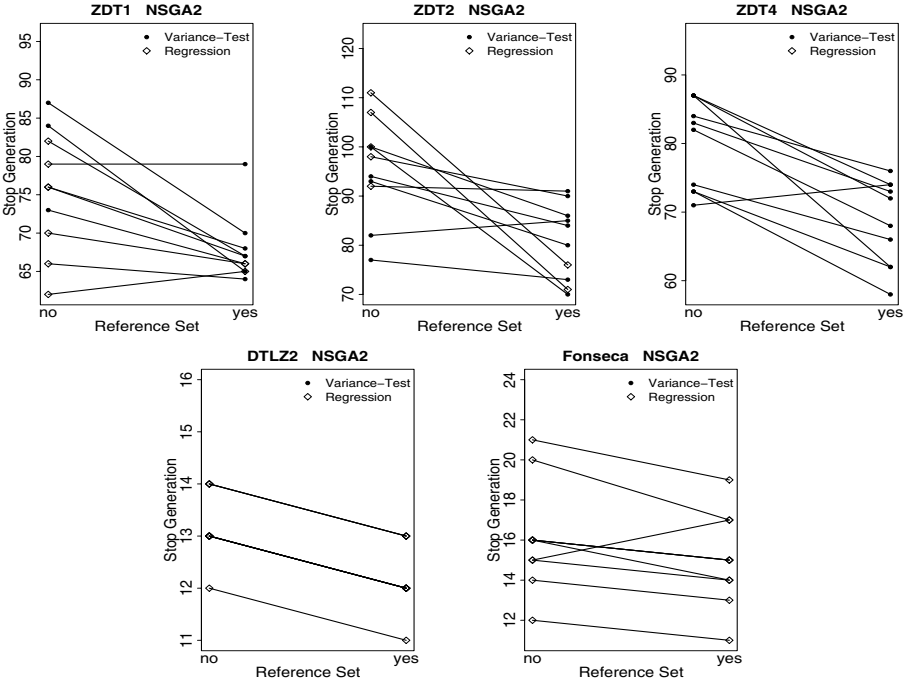

Fig. 6. In the line plots, the generations of NSGA2, in which the OCD stopping criterion is first met (left), are connected to the corresponding generations of OCD with full information (right). Furthermore, the test, which initiates the termination, is indicated by a specific symbol.

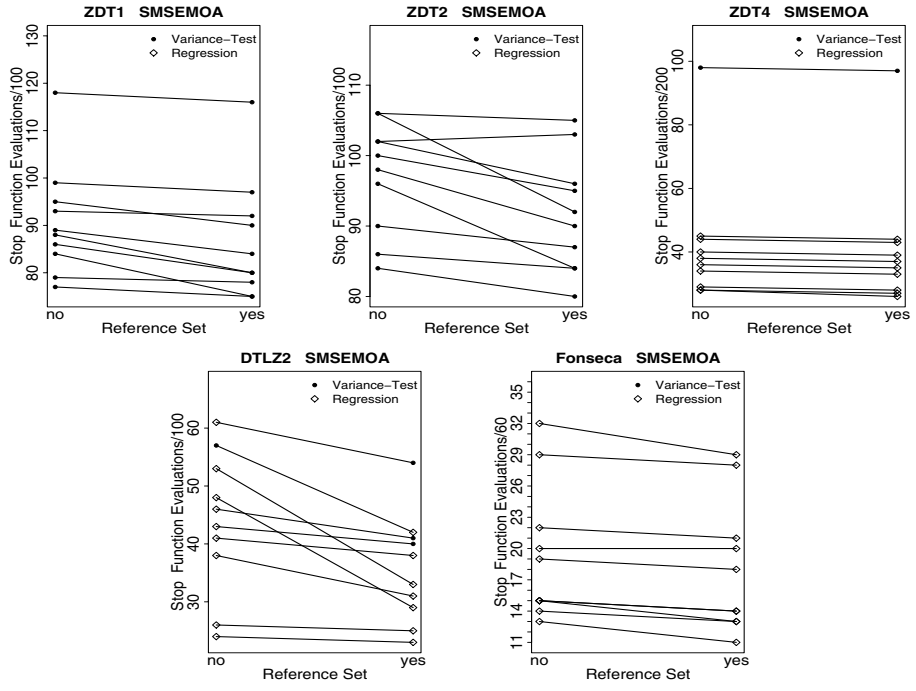

Fig. 7. In the line plots, the generations of SMSEMOA, in which the OCD stopping criterion is first met (left), are connected to the corresponding generations of OCD with full information (right). Furthermore, the test, which initiates the termination, is indicated by a specific symbol.

with extreme differences can be detected (cf. Fig. 6). Nevertheless, the generations proposed by OCD are matching the subjective localization of the termination generation with an accuracy of approximately $nPreGen = 10$ generations.

Discussion. The χ^2-variance test as well as the test on the regression coefficient are necessary to successfully detect convergence of EMOA. While the former indicates a low level of improvement in cases of successful optimization, e.g., on ZDT1 and ZDT2, the latter is extremely important when the high variance in the indicator values does not provide further improvements due to cyclic deterioration effects. These effects can be observed for NSGA2 on DTLZ2 and Fonseca. In contrast, on ZDT4 phases of temporary stagnation lead to the termination of the SMSEMOA based on the regression criterion. Due to a lower selection pressure, NSGA2 can avoid these phases and is therefore stopped by the variance criterion after global convergence.

Another important observation is that, in cases, in which OCD terminates the EMOA based on the χ^2-variance test, the value of $\sqrt{VarLimit} = 0.001$ is close to the differences in approximation quality compared to the one after the commonly proposed $MaxGen2$ FE. Thus, the user can approximately adjust the desired level of approximation accuracy ϵ by choosing $VarLimit = \epsilon^2$. However, the figures show that the value $VarLimit = 0.001^2$ is suitable for the considered test cases.

Table 2. Summary of $\boldsymbol{PI}_{i,j}$ and generation differences at the stop generation of OCD denoted as OCDStop and $MaxGen2$, where PIDiff = $\boldsymbol{PI}_{j,OCDStop} - \boldsymbol{PI}_{j,MaxGen2}$ and GenDiff = $MaxGen2-$OCDStop (j = {HV, EPS, R2}). Additionally, the number of saved function evaluations and their percentage of MaxGen2 are calculated.

		NSGA2		SMSEMOA	
problem	PI	med(PIDiff)	med(GenDiff)	med(PIDiff)	med(GenDiff)
ZDT1	HV	2.07e-03	124	0.96e-03	112
	Eps	2.08e-03	12400 FE	1.31e-03	11200 FE
	R2	0.93e-03	62%	0.38e-03	56%
ZDT2	HV	2.56e-03	104	1.01e-03	101
	Eps	3.13e-03	10400 FE	0.91e-03	10100 FE
	R2	1.46e-03	52%	0.63e-03	51%
ZDT4	HV	0.26e-03	18	21.72e-03	63
	Eps	0.28e-03	3600 FE	19.75e-03	12600 FE
	R2	0.06e-03	18%	9.07e-03	63%
DTLZ2	HV	-0.39e-03	287	0.72e-03	256
	Eps	-14.76e-03	28700 FE	3.37e-03	25600 FE
	R2	0.06e-03	96%	0.02e-03	85%
Fonseca	HV	0.97e-03	50	2.49e-03	49
	Eps	-0.51e-03	3000 FE	5.14e-03	2940 FE
	R2	0.14e-03	76%	0.21e-03	74%

The experiments document the general ability of the statistical tests within OCD to detect convergence based on performance indicator values. The delayed detection of convergence on the Fonseca problem is due to the time window of preceding generations and the very fast convergence of the EMOA. For a faster detection of stagnation, $nPreGen$ has to be decreased. However, the time of convergence as indicated by OCD can be accounted as premature for SMSEMOA on ZDT4 and DTLZ2 regarding the run of the metrics in further generations. In such situations, a larger time window allows longer phases of stagnation and provides the EMOA with the possibility to escape from local optima. In summary, a conflict between a fast detection of convergence and robustness with respect to short phases of stagnation exists. Therefore, the specification of the length of the time windows $nPreGen$ allows the user of OCD to express his own preferences based on the expected kind of problem.

The problem of the overestimation of the generation, in which stagnation occurs, when OCD is applied within NSGA2 can be explained by the selection that is implemented within this EMOA. Due to the high number of non-dominated solutions in the already converged population, the individuals are mainly evaluated by means of the crowding distance [16]. Thus, in combination with the $(\mu + \mu)$ selection, the population is still in motion. Since the reference set itself is part of this motion, a high variance in the indicator values is likely to appear. In contrast, SMSEMOA does only accept solutions, which increase the hypervolume of the current population. Thereby, a monotonic improvement can be expected, which also guarantees appropriate reference sets for OCD.

5 Conclusion

In this paper, a robust and reliable method for convergence detection within evolutionary multi-objective optimization algorithms has been introduced. This method is based on two statistical tests, namely the t-test on the regression coefficient and the χ^2-variance test, which guarantee an accurate convergence detection in all the considered examples. The proposed method is able to invoke different performance indicators, and it was investigated using the three recommended metrics from the EMO field. This way, we have been able to save half of the function evaluations for common test cases without having to accept a considerable loss of quality. However, the application of OCD to optimization scenarios, which include temporary phases of stagnation, such as in discrete optimization, could result in a premature indication of convergence.

In addition, we tried OCD on an already solved practical example [33], which is not shown due to a lack of space. This test indicated that the former analysis wasted many computational resources. Processing this hint by means of comprehensive evaluations of OCD on real-world problems is a task for the near future.

Furthermore, the technique of OCD offers a way for algorithm comparison. For this purpose, all EMOA parameters and operators have to be set to comparable values, and a high number of parallel runs of each benchmarked EMOA has to be performed. This way, a proper statistical analysis on the distributions of the stop generations proposed by OCD combined with the internally used performance indicators becomes possible. In this context, a comparison to an approach for offline convergence detection, which has been recently proposed by one of the authors [18], seems revealing.

Acknowledgements. This paper is based on investigations of the collaborative research center SFB/TR TRR 30, which is kindly supported by the *Deutsche Forschungsgemeinschaft (DFG)*. Moreover, we acknowledge financial support from the German *Federal Ministry of Economics and Technology (BMWi)*.

References

1. Zitzler, E., Künzli, S.: Indicator-based selection in multiobjective search. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiňo, P., Kabán, A., Schwefel, H.-P. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 832–842. Springer, Heidelberg (2004)
2. Beume, N., Naujoks, B., Emmerich, M.: SMS-EMOA: Multiobjective selection based on dominated hypervolume. European Journal of Operational Research 181(3), 1653–1669 (2007)
3. Zitzler, E., Thiele, L., Bader, J.: SPAM: Set preference algorithm for multiobjective optimization. In: Rudolph, G., Jansen, T., Lucas, S., Poloni, C., Beume, N. (eds.) PPSN 2008. LNCS, vol. 5199, pp. 847–858. Springer, Heidelberg (2008)
4. Wagner, T., Beume, N., Naujoks, B.: Pareto-, aggregation-, and indicator-based methods in many-objective optimization. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds.) EMO 2007. LNCS, vol. 4403, pp. 742–756. Springer, Heidelberg (2007)

5. Emmerich, M., Giannakoglou, K., Naujoks, B.: Single- and multi-objective evolutionary optimization assisted by gaussian random field metamodels. IEEE Trans. on Evolutionary Computation 10(4), 421–439 (2006)
6. Knowles, J.: ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE Trans. on Evolutionary Computation 10(1), 50–66 (2006)
7. Ponweiser, W., Wagner, T., Biermann, D., Vincze, M.: Multiobjective optimization on a limited amount of evaluations using model-assisted $\mathcal{S}$-metric selection. In: Rudolph, G., Jansen, T., Lucas, S., Poloni, C., Beume, N. (eds.) PPSN 2008. LNCS, vol. 5199, pp. 784–794. Springer, Heidelberg (2008)
8. Deb, K., Mohan, M., Mishra, S.: A fast multi-objective evolutionary algorithm for finding well-spread pareto-optimal solutions. KanGAL report 2003002, Indian Institute of Technology, Kanpur, India (2003)
9. Huang, V.L., Qin, A.K., Deb, K., Zitzler, E., Suganthan, P.N., Liang, J.J., Preuss, M., Huband, S.: Problem definitions for performance assessment of multi-objective optimization algorithms. Technical report, Nanyang Technological University (2007)
10. Rudolph, G., Naujoks, B., Preuss, M.: Capabilities of EMOA to detect and preserve equivalent pareto subsets. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds.) EMO 2007. LNCS, vol. 4403, pp. 36–50. Springer, Heidelberg (2007)
11. Kumar, A., Sharma, D., Deb, K.: A hybrid multi-objective optimization procedure using PCX based NSGA-II and sequential quadratic programming. In: Michalewicz, Z., Reynolds, R.G. (eds.) Congress on Evolutionary Computation (CEC). IEEE Press, Piscataway (2007)
12. Deb, K., Lele, S., Datta, R.: A hybrid evolutionary multi-objective and SQP based procedure for constrained optimization. In: Kang, L., Liu, Y., Zeng, S. (eds.) ISICA 2007. LNCS, vol. 4683, pp. 36–45. Springer, Heidelberg (2007)
13. Emmerich, M., Deutz, A., Beume, N.: Gradient-based/Evolutionary relay hybrid for computing pareto front approximations maximizing the S-metric. In: Bartz-Beielstein, T., Blesa Aguilera, M.J., Blum, C., Naujoks, B., Roli, A., Rudolph, G., Sampels, M. (eds.) HCI/ICCV 2007. LNCS, vol. 4771, pp. 140–156. Springer, Heidelberg (2007)
14. Deb, K., Jain, S.: Running performance metrics for evolutionary multi-objective optimization. In: Simulated Evolution and Learning (SEAL), pp. 13–20 (2002)
15. Rudenko, O., Schoenauer, M.: A steady performance stopping criterion for pareto-based evolutionary algorithms. In: Multi-Objective Programming and Goal Programming (2004)
16. Deb, K., Pratap, A., Agarwal, S.: A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans. on Evolutionary Computation 6(8) (2002)
17. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C., Fonseca, V.: Performance assessment of multiobjective optimizers: An analysis and review. IEEE Trans. on Evolutionary Computation 8(2), 117–132 (2003)
18. Trautmann, H., Ligges, U., Mehnen, J., Preuss, M.: A convergence criterion for multiobjective evolutionary algorithms based on systematic statistical testing. In: Rudolph, G., Jansen, T., Lucas, S., Poloni, C., Beume, N. (eds.) PPSN 2008. LNCS, vol. 5199, pp. 825–836. Springer, Heidelberg (2008)
19. Knowles, J., Thiele, L., Zitzler, E.: A tutorial on the performance assessment of stochastic multiobjective optimizers. 214, Computer Engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH) Zurich (2005)

20. Bleuler, S., Laumanns, M., Thiele, L., Zitzler, E.: PISA – A platform and programming language independent interface for search algorithms. In: Fonseca, C.M., Fleming, P.J., Zitzler, E., Deb, K., Thiele, L. (eds.) EMO 2003. LNCS, vol. 2632, pp. 494–508. Springer, Heidelberg (2003)
21. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, 2nd edn. Chapman and Hall, Boca Raton (2000)
22. Dudoit, S., van der Laan, M.: Multiple Testing Procedures with Applications to Genomics. Springer, Berlin (2008)
23. Stapleton, J.H.: Linear Statistical Models. Wiley Series in Probability and Statistics. Wiley, New York (1995)
24. Jensen, M.T.: Reducing the run-time complexity of multiobjective EAs: The NSGA-II and other algorithms. IEEE Trans. on Evolutionary Computation 7(5), 503–515 (2003)
25. Beume, N., Rudolph, G.: Faster S-metric calculation by considering dominated hypervolume as Klee's measure problem. In: International Conference on Computational Intelligence (CI 2006) (2006)
26. Ihaka, R., Gentleman, R.: R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5, 299–314 (1996)
27. Fonseca, C.M., Fleming, P.J.: Multiobjective genetic algorithms made easy: Selection, sharing, and mating restriction. In: Genetic Algorithms in Engineering Systems: Innovations and Applications, pp. 42–52 (1995)
28. Zitzler, E., Deb, K., Thiele, L.: Comparison of multiobjective evolutionary algorithms: Empirical results. Evolutionary Computation 8(2), 173–195 (2000)
29. Deb, K., Thiele, L., Laumanns, M., Zitzler, E.: Scalable multi-objective optimization test problems. In: Congress on Evolutionary Computation (CEC), vol. 1, pp. 825–830. IEEE Press, Piscataway (2002)
30. Deb, K.: Multi-objective Optimization using Evolutionary Algorithms. Wiley, Chichester (2001)
31. Zitzler, E., Thiele, L.: Multiobjective optimization using evolutionary algorithms - A comparative case study. In: Eiben, A.E., Bäck, T., Schoenauer, M., Schwefel, H.-P. (eds.) PPSN 1998. LNCS, vol. 1498, pp. 292–301. Springer, Heidelberg (1998)
32. Hansen, M.P., Jaszkiewicz, A.: Evaluating the quality of approximations to the non-dominated set. Technical Report IMM-REP-1998-7 (1998)
33. Wagner, T., Michelitsch, T., Sacharow, A.: On the design of optimisers for surface reconstruction. In: Thierens, D., et al. (eds.) 9th Annual Genetic and Evolutionary Computation Conference (GECCO 2007), Proc., London, UK, pp. 2195–2202. ACM, New York (2007)