On Metrics for Comparing Non-Dominated Sets

Joshua Knowles[1] and David Corne[2]
[1] IRIDIA, Free University of Brussels, Belgium, jknowles@ulb.ac.be
[2] Dept. of Computer Science, University of Reading, UK, d.w.corne@reading.ac.uk

December 20, 2001

Abstract: Evolutionary multi-objective optimisation (EMO) now boasts a proliferation of algorithms and benchmark problems. We need principled ways to compare the performance of different EMO algorithms, but this is complicated by the fact that the result of an EMO run is not a single scalar value, but a collection of vectors forming a non-dominated set. Various metrics for non-dominated sets have been suggested. Here we compare several, using the framework of 'outperformance relations' (Hansen and Jaszkiewicz [4]). This enables us to criticise and contrast a variety of published metrics, leading to some recommendations on which seem most useful in practice.

1 Introduction

Evolutionary multi-objective optimisation (EMO) is growing rapidly and several successful algorithms have emerged recently [6, 10, 3, 1]. The proliferation of such algorithms, along with a growing interest in benchmark multiobjective problems [10, 2, 8]) boosts the importance of performance comparison issues. This is complicated, however, since the result of an EMO run is not a single scalar 'best-fitness' value which can be subjected to univariate statistical tests. Instead, the result (except in pathological cases) is a non-dominated collection of vectors (given any two vectors in such a set that are not equal on all objectives, one cannot be better or equal to the other on all objectives – i.e. no vector in the set dominates another).

In the absence of other factors (e.g. preference for certain objectives, or for a particular region of the tradeoff surface), the task of an EMO algorithm is to yield as good an approximation as it can to the true Pareto front. Comparing two EMO algorithms hence requires comparing the nondominated sets they produce. Pathological cases aside, there is no straightforward way to distinguish the quality of different non-dominated sets. E.g. if we call the true Pareto front Z^*, then how do we compare a result which produces a single point $a \in Z^*$ with another result which yields a widespread set of non-dominated points B, none of which is either dominated by a or in Z^*? In the former case we have a true Pareto point, which means that no solution exists which dominates the discovered solution. But in the latter case, although every discovered point *can* be dominated, we have perhaps a very wide representation of the shape of the tradeoff surface (absent in the former case) and perhaps some of the points are not very far from true Pareto optimal.

Until recently, it has been common in EMO literature for performance to be indicated simply by way of a graphic plot. As Van Veldhuizen [9] remarks (but using our notation): "Comparative results are then 'clearly' shown in graphical form indicating which algorithm performed better, often implying the new MOEA's returned Z is a better representation of Z^*." More recently, several metrics have been proposed for comparing non-dominated sets, each of which attempts to boil down the (maybe comparative) 'quality' of such a set into a single number. Such a measure can then underpin statistical comparisons between EMO algorithms as well as convergence studies.

In the remainder we discuss and contrast several recently proposed non-dominated set comparison metrics (NDSCMs), in the light of the goals of EMO search, and also in the light of a framework proposed by Hansen and Jaszkiewicz ([4]) which enables us to compare them in terms of the kind of ordering an NDSCM induces on non-dominated sets. In section 2 we review the 'goals' of EMO search, and also set out the framework for assessing NDSCMs proposed by Hansen and Jaszkiewicz [4]. In Section 3 we analyse several NDSCMs from the recent literature with reference to the framework described in section 2, leading to a brief summary and recommendations in section 4.

2 EMO Objectives and Outperformance Relations

To properly compare and contrast NDSCMs we need to identify desirable aspects of non-dominated sets. Zitzler et al [11] suggest three goals that can be identified and measured:

1. The distance of the resulting nondominated set to the Pareto-optimal front should be minimized.
2. A good (in most cases uniform) distribution of the solutions found is desirable.
3. The extent of the obtained nondominated front should be maximized, i.e., for each objective, a wide range of values should be present.

These describe desirable outcomes, but we can question whether they fully capture EMO comparison needs. E.g. if $|Z^*| = 1$, then (3) is not appropriate. Also, if points on Z^* are not uniformly distributed, then a result Z that contains nearly all of the points in Z^* will not comply with (2). Thus, although the above serve well as an intuitive guide to the goals of Pareto search, a more general (and economic) statement might be to

expand on the first point only, defining what is meant by the distance of one set from the other. This may lead to more useful metrics.

More recently, Hansen and Jaszkiewicz [4] have considered the problem of evaluating approximations to the true Pareto front. They define a number of *outperformance relations* that express the relationship between two sets of internally nondominated objective vectors, A and B, as follows. where $\mathsf{ND}(S)$ denotes the non-dominated points in S:

Weak Outperformance: $AO_W B \iff \mathsf{ND}(A \cup B) = A$ and $A \neq B$. I.e. *A weakly outperforms B* if all points in B are 'covered' by those in A (where 'covered' means is equal to or dominates) and there is at least one point in A that is not contained in B.

Strong outperformance: $AO_S B \iff \mathsf{ND}(A \cup B) = A$ and $B \setminus \mathsf{ND}(A \cup B) \neq \emptyset$. I.e. *A strongly outperforms B* if all points in B are covered by those in A and some point in B is dominated by a point in A.

Complete outperformance: $AO_C B \iff \mathsf{ND}(A \cup B) = A$ and $B \cap \mathsf{ND}(A \cup B) = \emptyset$. I.e. *A completely outperforms B* if each point in B is dominated by a point in A.

Notice that $AO_C B \Rightarrow AO_S B \Rightarrow AO_W B$. In other words, complete outperformance is the strongest and weak outperformance is the weakest of the relations.

These valuably describe the relationships between approximations to Z^* since they are compatible with, and only depend upon, standard Pareto dominance. They are not metrics of performance, however, and are silent in the often encountered case where each set contains points that are not covered by the other set. But, we can use them to assess the usefulness of NDSCMs. Any metric not compatible with these relations is lcapable of giving misleading results. In this vein, Hansen and Jaszkiewicz formally define compatibility and weak compatibility with an outperformance relation, as follows:

Weak compatibility: A comparison metric R is *weakly compatible* with an outperformance relation $\leq$ if for each pair of nondominated sets A, B with $A \leq B$, R will evaluate A as being no worse than B.

Compatibility: A comparison metric R is *compatible* with an outperformance relation $\leq$ if for each pair of nondominated sets A and B, such that $A \leq B$, R will evaluate A as being better than B.

In the remainder, we compare and contrast different proposed NDSCMs in terms of their compatibility with the outperformance relations.

3 Analysis of NDSCMs

In this section we analyse several NDSCMs, discussing their compatibility with the outperformance relations and other relevant factors. For each metric, we introduce it and then discuss it under three headings: **Pareto compatibility**, **Pros**, **Cons/caveats**. The analysis is summarized in Table 1 When introduced, we note its purpose and how it actually compares two approximation sets A and B. There are several alternative approaches to this. A 'direct comparative' metric compares A and B directly using a scalar measure $R(A, B)$ to describe how much better A is than B. If $R(A, B) = c - R(B, A)$ for some constant c for all pairs of nondominated sets A, B then R is 'symmetric'. Alternatively, a 'reference metric' uses a reference set, perhaps Z^*; it scores both sets against this reference set, and then compares the results. Clearly, any direct comparative metric can also be used as reference metric by specifying a particular reference set. The converse is not true because their definition depends on a particular reference set (often Z^*). Lastly, an 'independent' metric measures some property of each set that is not dependent on any other, or any reference set. Another important feature of a metric is whether it induces a complete ordering of all possible nondominated sets. This ensures transitivity, so that when A, B, and C are compared, if A beats B and B beats C then it is always true that A beats C. Often, direct comparative metrics do not induce a complete ordering, and the relations between different sets may be intransitive. Using reference sets in such cases would then ensure transitivity. Transitivity is not generally a problem with independent metrics as they all induce a complete ordering. Finally, we also note if the metric is a cardinal measure (based on counting the number of vectors in some set) or non-cardinal.

The Pareto compatibility section of each analysis is concerned with compatibility with the outperformance relations O_W, O_S, and O_C. The less compatible the metric is, the more misleading it may be, giving scores for nondominated sets that do not accurately reflect their relative worth in a Pareto sense. The hardest relation to be (weakly) compatible with is O_W, and the easiest is O_C. We note that compatibility with O_W is necessary and sufficient for ensuring **monotony** and sufficient but not necessary for ensuring **relativity**, which are defined as follows (and are clearly desirable features of an NDSCM):

(weak) monotony Given a nondominated set A, adding a nondominated point improves (does not degrade) its evaluation.

(weak) relativity The evaluation of Z^* is (non)-uniquely optimal, i.e., all other nondominated sets have a strictly inferior (non-superior) evaluation.

Weak compatibility with O_W is sufficient for the weak versions to be exhibited.

The last two headings summarize the advantages and disadvantages of the metric, considering compatibility with the outperformance relations and additional factors such as computational cost, whether or not it is scaling independent (is the ordering of approximations affected if one objective is scaled relative to the others?), and whether it relies on knowledge of Z^* or any other reference set or point, and whether it can differentiate between different levels of complete outperformance. This means that given three approximation sets A, B, C with $AO_C B$ and $BO_C C$, would the metric give a different evaluation if A and B were compared than if A and C were compared?

The $\mathcal{S}$ metric

A definition of the $\mathcal{S}$ metric is given in [10]. It calculates the hypervolume of the multi-dimensonal region

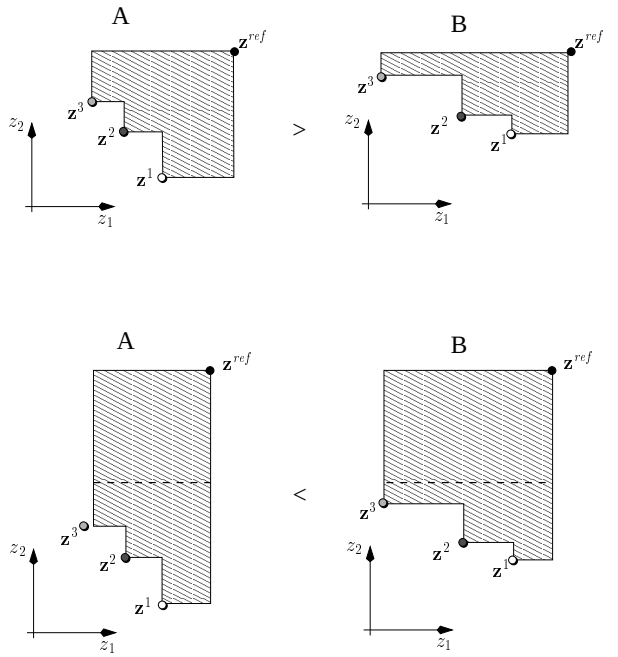

Figure 1: The relative value of the $\mathcal{S}$ metric depends upon an arbitrary choice of reference point. In the upper half, two nondominated sets are shown, A and B, with $\mathcal{S}(A) > \mathcal{S}(B)$. In the lower half the same sets have a different ordering in $\mathcal{S}$.

enclosed by A and a 'reference point' (see Figure 1), hence computing the size of the region A dominates. It is independent (although needs a reference point to be chosen), so it induces a complete ordering, and it is non-cardinal. **Pareto compatibility**: Compatible with O_W provided that the reference point is set so that all feasible nondominated sets are evaluated as positive. **Pros**: There are many: compatible with the outperformance relations, independent, differentiates between different degrees of complete outperformance of two sets, scaling independent, and its meaning is intuitive. **Cons/caveats**: It requires defining some upper boundary of the region within which all feasible points will lie. This choice does affect the ordering of nondominated sets (see Figure 1), and is relatively arbitrary. It has a large computational overhead, $O(n^{k+1})$, rendering it unusable for many objectives or large sets. It multiplies 'apples' by 'oranges', that is, different objectives together, but arguably this does not matter, since the metric is scaling independent anyway, and the units are irrelevant.

Error ratio (ER)

This [8] is defined as $(\sum_{i=1}^{n} e_i)/n$, where n is the number of vectors in the approximation set Z; $e_i = 0$ if vector i is in Z^* and 1 otherwise. Lower values of the error ratio represent better nondominated sets. ER is the proportion of non true Pareto points in Z. It is a reference metric using Z^* as reference set. It induces a total ordering, and is cardinal. **Pareto compatibility**: It is only weakly compatible with O_C. It is not weakly compatible with O_S or O_W; e.g. if an algorithm finds two nondominated vectors, one in Z^*, and the other far from Z^*, then its error ratio is 0.5. If it finds one hundred solutions, 99 of which are very close to Z^* (and perhaps distributed evenly along it over a wide range in the objectives), and one (as before) which is in Z^*, then its error ratio will be 0.99. Clearly the second set of points is far better, but the first has a much worse ER. It strongly violates monotony; given a nondominated set A with one or more Pareto optimal points in it, addition of more nondominated but non-Pareto optimal points makes the ER score worse. It violates relativity too, since any non-empty subset of Z^* has an opimal error ratio. However, it exhibits weak relativity because the Pareto front itself is evaluated not worse than any other set. **Pros**: It is easy to understand and easy to calculate. It is scaling independent. For test problems it can be used as a quick and rough means of assessing progress towards Z^*. **Cons/caveats**: Knowledge of Z^* is needed. It is incompatible with the outperformance relations.

Generational distance (GD)

This [8] is $\frac{\sqrt{\sum_{i=1}^{n} d_i^2}}{n}$, where n is the number of vectors in the approximation set, and d_i is the distance in objective space between vector i and the *nearest* member of Z^*. Lower values represent better sets. GD measures general progress towards Z^*. It is a reference metric using Z^* as reference. It induces a total ordering, and is non-cardinal. **Pareto compatibility**: It is not weakly compatible with O_W, but is compatible with O_S. It violates weak monotony. E.g., the GD score favours one vector close to Z^* over a set containing that vector plus others, as long as the others are not closer on average to Z^* than the first one. It does exhibit weak relativity, since any subset of Z^* has an optimal GD. **Pros**: For a constant size of nondominated set, GD is compatible with O_S. It is relatively cheap to calculate. **Cons/caveats**: Because it is not compatible with O_W it cannot be used confidently for nondominated sets that are changing in cardinality (typical, for example, of the non-dominated portion of an EMO population over time). It cannot reliably differentiate between different levels of complete outperformance. Knowledge of Z^* is required. The distance metric will either add or multiply different objectives together, introducing scaling and normalization issues that cannot be properly resolved without reference to additional preference information.

Maximum Pareto Front Error (MPFE)

This [8] is defined as follows:

$$\max_j(\min_i |f_1^i(\vec{x}) - f_1^j(\vec{x})|^p + |f_2^i(\vec{x}) - f_2^j\ (\vec{x})|^p)^{1/p} \quad (1)$$

where $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$ index vectors in Z and Z^* respectively, and $p = 2$. Lower values represent better sets. MPFE measures the largest distance between any vector in Z and the corresponding closest vector in Z^*. It is a reference metric using Z^* as a reference. It induces a complete ordering, and is non-cardinal. **Pareto compatibility** It is not weakly compatible with any outperformance relation. It violates weak monotony. It is better, according to MPFE,

to find one solution close to Z^* than to find ten solutions, nine of which are in Z^* and one which is some distance away. This does not sit well with typical intuitions about the quality of a nondominated set. It exhibits **weak relativity** because any subset of Z^* is optimal. **Pros**: It is cheap to compute. It provides information about whether any points found are far from the true front. **Cons/caveats** Even if a nondominated set has a very low MPFE, it does not make it a good front, and doesn't necessarily make it better than another one with a much worse MPFE. As with the other distance metrics, different objectives must be combined to get a single figure of merit, bringing in scaling and normalization issues. Knowledge of Z^* is required.

Overall Nondom. Vector Generation (ONVG)

A further metric in [8], ONVG, is simply defined as $|Z|$. Measuring the number of distinct non-dominated points produced, this is an independent metric, induces a complete ordering on the set of approximations, and it is a cardinal measure. **Pareto compatibility**: It is not weakly compatible with any outperformance relation. It does not exhibit either weak monotony or weak relativity. **Pros**: It is easy to compute. It is scaling independent. There are a few special pathological cases where this metric can be used to gauge the quality of a nondominated set, for example, if the entire search space contains only nondominated points. **Cons/caveats**: See Pareto compatibility. In general, it is straightforward to come up with scenarios in which A outperforms B on this metric but in which B is clearly 'better' than A. E.g. A contains a million nondominated points and B contains just 1, but this point dominates all of those in A.

Van Veldhuizen ([8]) also defines 'Overall nondominated vector generation ratio' (ONVGR) as $|Z|/|Z^*|$. We omit a fuller discussion of this and further metrics proposed in [8], since the above are representative, and fuller analysis can be found in [5].

Schott's Spacing metric (SS)

Schott [7] describes the following spacing metric:

$$\sqrt{\frac{1}{n-1}\sum_{i+1} n(\bar{d} - d_i)^2} \qquad (2)$$

where $d_i = min_j(|f_1^i(\vec{x}) - f_1^j(\vec{x})| + |f_2^i(\vec{x}) - f_2^j(\ \vec{x}|))$, $i, j = 1..n$, $\bar{d}$ is the mean of all d_i and $n = |Z|$.

SS tries to gauge how evenly the points are distributed. It is independent metric, induces a complete ordering, and is cardinal. **Pareto compatibility**: SS is not even weakly compatible with O_W. It exhibits nether monotony nor relativity, since Z^* may be non-uniform. **Pros**: Used in conjunction with other metrics (as it is designed to be), it provides information about the distribution of vectors obtained. It has low computational overhead. It can be generalized to more than two dimensions by extending the definition of d_i. **Cons/caveats**: Schott's definition of d_i does not specify the use of normalized distances, which may be problematic. Its incompatibility with the outperformance relations and the fact that it violates both monotony and relativity make it unreliable.

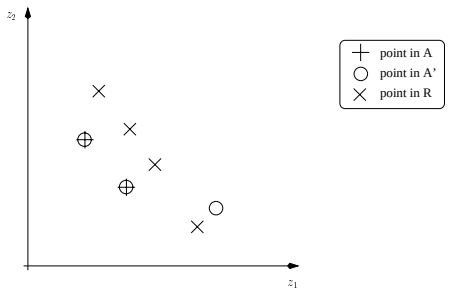

Figure 2: $\mathcal{C}(R, A) = 0$, $\mathcal{C}(R, A') = 1/3$, $\mathcal{C}(A, R) = 3/4$, $\mathcal{C}(A', R) = 3/4$ so against the reference set R, A' *seems to* evaluate worse than A according to the $\mathcal{C}$ metric. But $A' O_W A$ so we may conclude that the $\mathcal{C}$ metric is not weakly compatible with the weak outperformance relation when used with a general reference set R. The alternative view is to say that A is not evaluated 'worse' than A by the $\mathcal{C}$ metric, arguing that we should adopt the convention that unless one of a $\mathcal{C}$ metric pair gives output 1, the evaluation is meaningless. However, that severely restricts its general usefulness.

Deb et al [3] also define a spacing metric. We omit details for space reasons, but can say that it turns out to have very similar properties to Schott's metric in the sense of table 1 [5].

The $\mathcal{C}$ metric

Let $A, B \subseteq X$ be two sets of vectors. $\mathcal{C}$ maps the ordered pair (A, B) to the interval [0,1]:

$$\mathcal{C}(A, B) = \frac{|\{b \in B | \exists a \in A : a \preceq b\}|}{|B|}$$

The value $\mathcal{C}(A, B) = 1$ means that all decision vectors in B are weakly dominated by A. The opposite, $\mathcal{C}(A, B) = 0$, represents the situation when none of the points in B is weakly dominated by A. Note that always both orderings have to be considered, since $\mathcal{C}(A, B)$ is not necessarily equal to $1 - \mathcal{C}(B, A)$.

$\mathcal{C}$ is a cardinal measure, and a direct comparative approach giving a single figure of merit that is not symmetric. It is difficult to establish whether the metric induces a complete ordering because it is not clear how the pair of $\mathcal{C}$ values should be interpreted together.

Pareto compatibility: The non-symmetric nature of $\mathcal{C}$ complicates the analysis of its compatibility with the outperformance relations. This depends on how we interpret or combine the two outputs of the metric. Space restrictions force us to omit a fuller discussion, which can be found in [5]; here we provide a summary.

It is generally possinble for $\mathcal{C}(A, B)$ to differ from $\mathcal{C}(B, A)$. However, if we take it that in general a set A is evaluated better than a set B according to $\mathcal{C}$ if $\mathcal{C}(A, B) = 1$ and $\mathcal{C}(B, A) < 1$, then $\mathcal{C}$ is compatible with O_W. There are further constrained situations in which $\mathcal{C}$ can be coerced into compatibility with O_W, when used either as a direct comparative metric or as a reference set metric. However, in general (fuller dis-

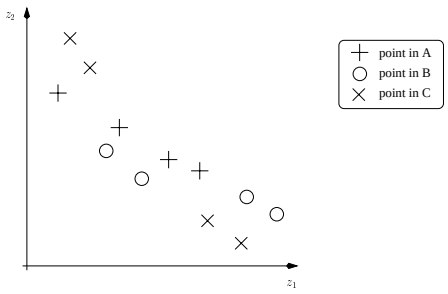

Figure 3: Cycling in the $\mathcal{C}$ metric. $\mathcal{C}(A,B) = 0, \mathcal{C}(B,A) = 3/4, \mathcal{C}(B,C) = 0, \mathcal{C}(C,B) = 1/2$ and $\mathcal{C}(A,C) = 1/2, \mathcal{C}(C,A) = 0$ so, $\mathcal{C}$ considers B better than A, C better than B, but A better than C

cussion in [5]), $\mathcal{C}$ is not compatible with O_W (see figure 2). It is, however, compatible with O_S and O_C.

Any pair of $\mathcal{C}$ metric scores for a pair of sets A and B in which neither $\mathcal{C}(A,B) = 1$ nor $\mathcal{C}(B,A) = 1$, indicates that the two sets are incomparable according to the weak outperformance relation. Drawing any further conclusions from $\mathcal{C}$ in this case is inadvisable. E.g. Figure 3 shows that if three sets are compared using $\mathcal{C}$, they may not be ordered. I.e. $\mathcal{C}$ is cycle-inducing. Further, $\mathcal{C}$ does not give an output which is representative of our intuitions about the relative quality of two sets *unless* the two sets contain very evenly distributed points, and are of very similar cardinality. **Pros**: It has low computational overhead compared to $\mathcal{S}$. It is compatible with O_S, and scale and reference point independent. It requires no knowledge of Z^*. For two evenly-distributed sets, of the same cardinality, it gives results compatible with intuitive notions of quality. **Cons/caveats**: Its incompatibility with O_W. If two sets are of different cardinality and/or the distributions of the sets are non-uniform, then it gives unreliable results. It cannot determine the degree of outperformance if one set completely outperforms the other.

$D1_R$ (Czyzak and Jaszkiewicz)

$$D1_R(A,\Lambda) = \frac{1}{|R|}\sum_{r\in R}\min_{\mathbf{z}\in A}\{d(\mathbf{r},\mathbf{z})\} \qquad (3)$$

where A is the approximation set, R is a reference set, $d(\mathbf{r},\mathbf{z}) = \max_k\{\lambda_k(r_k - z_k)\}$ and $\Lambda = [\lambda_1, \lambda_2, \ldots \lambda_K], \lambda_k = 1/R_k, k = 1..K$ with R_k being the range of objective k in set R.

$D1_R$ measures the mean distance, over the points in a reference set, of the nearest point in an approximation set. It is a reference metric which induces a complete ordering, and is non-cardinal. **Pareto compatibility**: $D1_R$ is weakly compatible with O_W, but not compatible with O_C. **Pros**: It is cheap to compute. Its weak compatibility with the outperformance relations. It *can* differentiate between different levels of complete outperformance given an appropriate choice of reference set. **Cons/caveats**: It effectively calculates a weighted average where the reference points have equal weight. Hence the score is strongly dependent upon the distribution of points in the reference set, and on the choice of Λ.

$R1$ and $R1_R$ (Hansen and Jaszkiewicz)

$$R1(A,B,U,p) = \int_{u\in U} C(A,B,u)p(u)du \text{ , where}$$

$$C(A,B,u) = \begin{cases} 1 & \text{if } u^*(A) > u^*(B) \\ 1/2 & \text{if } u^*(A) = u^*(B) \\ 0 & \text{if } u^*(A) < u^*(B) \end{cases}$$

where A and B are two approximation sets, U is some set of utility functions, $u : \Re^K \rightarrow \Re$ which maps each point in the objective space into a measure of utility, $p(u)$ is an intensity function expressing the probability density of the utility $u \in U$, and $u^*(A) = \max_{\mathbf{z}\in A}\{u(\mathbf{z})\}$ and similarly for $u^*(B)$.

$R1$ is based on calculating the probability that A is better than B over a set of utility functions. It is a direct comparative metric and non-cardinal. It does not induce a total ordering. $R1_R$ is $R1$ when it is used with a reference set, in which case it does induce a total ordering. **Pareto compatibility**: Assuming we are maximizing all objectives, a utility function u is strictly compatible with the dominance relation iff $\forall \mathbf{z}^1, \mathbf{z}^2 : \mathbf{z}^1 > \mathbf{z}^2 \Rightarrow u(\mathbf{z}^1) > u(\mathbf{z}^2)$. The set of all utility functions that are strictly compatible with the dominance relation is U_{sc}. Let $U(A > B) = \{u \in U \mid u^*(A) > u^*(B)\}$. If the probability density function $p(u)$ is such that the probability of selecting a utility function $u \in U(A > B)$ is positive whenever $U(A > B) \neq \emptyset$ and $U \subseteq U_{sc}$ then $R1$ is compatible with O_W. Under the same conditions, $R1_R$ is only weakly compatible with O_W and is not compatible with O_C. **Pros**: The metrics are scaling independent, and have a lower computational overhead than $\mathcal{S}$. $R1_R$ can differentiate between different levels of complete outperformance provided that an appropriate reference set is chosen. **Cons/caveats**: $R1$ is cycle-inducing. The metrics depend upon being able to define a set of utility functions. In general, however, this can be achieved without any knowledge of Z^*.

$R2$ and $R2_R$ (Hansen and Jaszkiewicz)

$$\begin{aligned} R2(A,B,U,p) &= E(u^*(A)) - E(u^*(B)) \\ &= \int_{u\in U}(u^*(A) - u^*(B))p(u)du \end{aligned}$$

where A and B are two approximation sets, U is some set of utility functions, $u : \Re^K \rightarrow \Re$ which maps each point in the objective space into a measure of utility, $p(u)$ is an intensity function expressing the probability density of the utility $u \in U$, and $u^*(A) = \max_{\mathbf{z}\in A}\{u(\mathbf{z})\}$ and similarly for $u^*(B)$.

Where $R1$ just uses the function $C(A,B,u)$ to decide which of two approximations is better on utility function u, without measuring by *how much*, $R2$ takes into account the expected values of the utility. $R2$ calculates the expected *difference* in the utility of an approximation A with another one B. It is a direct comparative metric. It induces a complete ordering. It is a non-cardinal measure. $R2_R$ is $R2$ when used as

a reference metric. It also induces a complete ordering. **Pareto compatibility**: $R2$ is compatible with O_W subject to the same set of conditions on the set of utility functions used as outlined for $R1$. $R2_R$ is also compatible with O_W given this set of conditions. **Pros**: The advantages of $R2$ arise from its compatibility with all of the outperformance relations and the fact that it can differentiate between different levels of complete outperformance. **Cons/caveats**: The application of $R2$ depends upon the assumption that it is meaningful to add the values of different utility functions from the set U. This simply means that that each utility function in U must be appropriately scaled with respect to the others and its relative importance.

$R3$ and $R3_R$ (Hansen and Jaszkiewicz)

Hansen and Jaszkiewicz also propose a similar metric to $R2$ whereby the *ratio* of the best utility values is calculated instead of the differences. These metrics are called $R3$ and $R3_R$. The latter is similar to the approach used in single objective optimization, where an approximate solution is evaluated by the ratio of its value to that of a fixed bound, note Hansen and Jaszkiewicz.

4 Summary of Analysis

From the analysis above, summarised in table 1, we would recommend to EMO researchers the use of the $R1$, $R2$, and $R3$ metrics of Hansen and Jaszkiewicz, and (for relatively few objective dimensions and not overlarge non-dominated sets) the $\mathcal{S}$ metric of Zitzler. The other metrics may not be as useful because they generally suffer from poor compatibility with the outperformance relations and cannot differentiate between different levels of complete outperformance.

$O_{W,S,C}$	C	R	S	c	D	$\mathcal{O}$	metric
2	$\times$	Y	Y	$\times$	$\times$	low	$\mathcal{C}$
3	Y	$\times$	Y	Y	Y	high	$\mathcal{S}$
2	Y	$\times$	$\times$	Y	$\times$	med	GD
0	Y	$\times$	Y	$\times$	$\times$	low	ER
0	Y	Y	$\times$	Y	$\times$	med	MPFE
0	Y	Y	Y	Y	$\times$	low	ONVG
0	Y	$\times$	Y	$\times$	$\times$	low	ONVGR
0	Y	Y	$\times$	Y	$\times$	med	SS
1	Y	$\times$	$\times$	Y	Y	med	$D1_R$
3	$\times$	$Y(\times)$	Y	Y	$\times$	med	$R1$
1	Y	$\times$	Y	Y	Y	med	$R1_R$
3	Y	$Y(\times)$	$\times$	Y	Y	med	$R2$
3	Y	$\times$	$\times$	Y	Y	med	$R2_R$

Table 1: Summary of the analysed metrics. Many Y's is generally good. Regarding compatibility with O_W, O_C, and O_S, the metrics fall into one of 4 groups. In column 1: "0" means not compatible with any of the three relations, "1" means weakly compatible with each of them, "2" means not compatible with O_W, but compatible with O_S and O_C, and "3" means compatible with all three. Y in the remaining columns indicates the following: C: non cycle-inducing; R: independent of reference set; S: independent of scaling (i.e. the ordering of approximations would not be affected by scaling the objectives differently); c: a non-cardinal measure. D: differentiates between levels of complete outperformance (see section 2). The $\mathcal{O}$ column indicuates computational overhead, and the final column indicates the metric under study in each row, using abbreviations indicated in the text.

Acknowledgments

Joshua Knowles thanks BT Labs Plc. David Corne thanks Evosolve Ltd for additional support.

References

[1] D. Corne, N. Jerram, J. Knowles, and M. Oates. PESA-II: Region-based selection in evolutionary multiobjective optimization. In *Proceedings of GECCO-2001: Genetic and Evolutionary Computation Conference*, pages 283–290. Morgan Kaufmann, 2001. (To appear).

[2] Kalyanmoy Deb. Multi-objective evolutionary algorithms: Introducing bias among pareto optimal solutions. Technical Report KanGAL 99002, Kanpur Genetic Algorithms Laboratory, Indian Institute of Technology Kanpur, Kanpur, India, 1999.

[3] Kalyanmoy Deb, Samir Agrawal, Amrit Pratab, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000.

[4] Michael Pilegaard Hansen and Andrzej Jaszkiewicz. Evaluating the quality of approximations to the non-dominate d set. Technical Report IMM-REP-1998-7, Technical University of Denmark, March 1998.

[5] Joshua Knowles. *Local-search and hybrid evolutionary algorithms for Pareto Optimization*. PhD thesis, University of Reading, Department of Computer Science, Reading, U.K., January 2002.

[6] Joshua D. Knowles and David W. Corne. Approximating the nondominated front using the Pareto archived evolution strategy. *Evolutionary Computation*, 8(2):149–172, 2000.

[7] Jason R. Schott. Fault tolerant design using single and multicriteria genetic algorithm optimization. Master's thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 1995.

[8] David A. Van Veldhuizen. *Multiobjective Evolutionary Algorithms: Classifications, Analyses, and New Innovations*. PhD thesis, Department of Electrical and Computer Engineering. Graduate School of Engineering. Air Force Institute of Technology, Wright-Patterson AFB, Ohio, May 1999.

[9] David A. Van Veldhuizen and Gary B. Lamont. Multiobjective evolutionary algorithms: Analyzing the state-of-the-art. *Evolutionary Computation*, 8(2):125–147, 2000.

[10] Eckart Zitzler. *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*. PhD thesis, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, November 1999.

[11] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation*, 8(2):173–195, Summer 2000.