Quality evaluation of solution sets in multiobjective optimisation

Li, Miqing; Yao, Xin

DOI:
10.1145/3300148

Document Version
Peer reviewed version

Citation for published version (Harvard):
Li, M & Yao, X 2019, 'Quality evaluation of solution sets in multiobjective optimisation: a survey', *ACM Computing Surveys*, vol. 52, no. 2, 26. https://doi.org/10.1145/3300148

Quality Evaluation of Solution Sets in Multiobjective Optimisation: A Survey

Miqing Li, and Xin Yao[1]

[1]CERCIA, School of Computer Science, University of Birmingham, Birmingham B15 2TT, U. K.
*Email: limitsing@gmail.com, x.yao@cs.bham.ac.uk

Abstract: Complexity and variety of modern multiobjective optimisation problems result in the emergence of numerous search techniques, from traditional mathematical programming to various randomised heuristics. A key issue raised consequently is how to evaluate and compare solution sets generated by these multiobjective search techniques. In this article, we provide a comprehensive review of solution set quality evaluation. Starting with an introduction of basic principles and concepts of set quality evaluation, the paper summarises and categorises 100 state-of-the-art quality indicators, with the focus on what quality aspects these indicators reflect. This is accompanied in each category by detailed descriptions of several representative indicators and in-depth analyses of their strengths and weaknesses. Furthermore, issues regarding attributes that indicators possess and properties that indicators are desirable to have are discussed, in the hope of motivating researchers to look into these important issues when designing quality indicators and of encouraging practitioners to bear these issues in mind when selecting/using quality indicators. Finally, future trends and potential research directions in the area are suggested, together with some guidelines on these directions.

Keywords: Quality evaluation, performance assessment, indicator, metric, measure, multobjective optimisation, multi-criteria optimisation, exact method, heuristic, metaheuristic, evolutionary algorithms

1 Introduction

In real world, it is not uncommon to face an optimisation problem with multiple objectives/criteria, namely, multiobjective optimisation problems (MOPs). These objectives are often conflicting, and there is no single optimal solution but instead a set of Pareto optimal solutions (termed a Pareto front in the objective space). A solution in the Pareto optimal set cannot be improved on an objective without degrading on some other objectives. For example, consider a car purchase problem where we want to buy a car with as good performance as possible but as low price as possible. Apparently, we cannot find a single car that achieves the best on both objectives. Figure 1 gives all types of cars (solutions) available in the market. As can be seen, there exist a set of solutions (black points), each of which is not inferior to any solution on both objectives. We may only be interested in these Pareto optimal solutions as the remaining solutions (white points) are always outperformed by some of them.

The growing interest in MOPs results in a variety of search techniques (called multiobjective optimisers hereinafter) Coello et al. (2007); Ehrgott (2006); Zhou et al. (2011), ranging from exact to approximation methods Brunsch and Röglin (2015); Miettinen (1999), and from heuristics to metaheuristics Czyzak and Jaszkiewicz (1998); Hansen (1997); Jones et al. (2002); Knowles and Corne (1999); Mandow and De La Cruz (2010). These multiobjective optimisers aim to generate

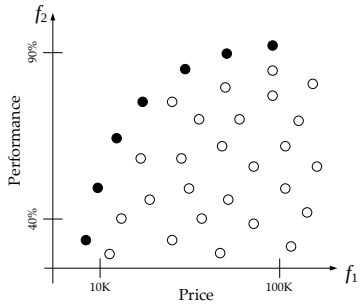

Figure 1: An example of the car purchase problem, where a user is interested in two conflicting objectives, price and performance. The points correspond to all types of cars (solutions) available in the market, in which the black ones correspond to Pareto optimal solutions and the white ones correspond to non-optimal solutions.

a solution set (aka a Pareto front approximation or an approximate set) that well represents the Pareto front of an MOP. Population-based non-numerical optimisers have been found promising in this regard, such as genetic algorithm Eiben and Smith (2015), particle swarm optimisation Coello et al. (2004) and ant colony optimisation Alaya et al. (2007). In such methods, each individual in the population aims to locate a unique trade-off among objectives, whereby they all together provide a representation of the whole Pareto front. A good Pareto front representation is of high importance in multiobjective optimisation. It presents a reliable and compact "picture" of the Pareto front and avoids overload in terms of the computation and the later decision-making process. From a representative solution set, the decision-maker can choose her/his favourable solution, or uses this information to articulate preference that allows to narrow down the search for a final choice.

An important issue in multiobjective optimisation is to evaluate and compare the quality of solution sets, e.g., those generated by different multiobjective optimisers. However, in the presence of multiple (or even an infinite number of) optimal solutions, two solution sets may be incomparable. This contrasts with single-objective optimisation, where we can typically define the quality of a solution set by its best solution (i.e., the solution with the smallest or largest objective function value). In multiobjective optimisation, the quality of a solution set usually contains several aspects, e.g., the closeness to the Pareto front, the coverage over the Pareto front, and the uniformity amongst solutions. This substantially complicates the comparison between solution sets.

A straightforward way to compare the quality of solution sets is visualisation. However, visual comparison fails to quantify the difference between solution sets and also becomes harder with more objectives involved. Quality indicators (QIs)[1], as a quantitative way of comparing sets, have arisen. They typically map a solution set to a real number that indicates one or several aspects of solution set quality, thus defining a total order amongst solution sets. QIs serve several purposes. They can be used to 1) compare competing multiobjective optimisers, reveal their strengths and weaknesses, and identify the most promising one; 2) monitor the search process of optimisers, optimise their parameters, and potentially improve their performance; 3) define the stop criterion of optimisers, especially those having a stochastic nature; and 4) explicitly guide the search by integrating into optimisers as selection criteria, e.g., instead of the Pareto dominance criterion.

Over the last two decades, quality evaluation of multiobjective solution sets has gained much attention, not only in the fields of evolutionary computation, operational research and optimisation Bozkurt et al. (2010); Sayın (2000); Zitzler et al. (2003), but also in other fields like artificial intelligence Bringmann and Friedrich (2013); Tan et al. (2002), software engineering Li et al. (2018a);

[1]There are other names of quality indicators in the literature, such as quality measures, performance indicators, and performance metrics.

Wang et al. (2016), and mechanical design Farhang-Mehr and Azarm (2003a); Wu and Azarm (2001). In general, studies on QIs in the literature can be classified into five categories.

1. Design of QIs. Majority of the QI studies lies in this category, aiming to design a reliable, efficient QI to evaluate and compare solution sets.

2. QI-based search. Coined by Zitzler and Künzli Zitzler and Künzli (2004), there is growing interest in the use of QIs to guide the search in multiobjective optimisation Bader and Zitzler (2011); Beume et al. (2007); Brockhoff et al. (2015); Jiang et al. (2015); Tian et al. (2017).

3. Theoretical studies of QIs. This category works on some important issues in the design of QIs, including analyses of the computational complexity of QIs Bringmann and Friedrich (2012); Fonseca et al. (2006); While et al. (2012), and discussions of some properties that a QI (or a combination of QIs) desires for, such as Pareto compatibility and completeness Knowles et al. (2006); Lizárraga-Lizárraga et al. (2008b); Zitzler et al. (2003), scaling invariance Zitzler et al. (2008), and minimum set construction Farhang-Mehr and Azarm (2003b).

4. Understanding of existing QIs. This includes the behaviour investigation of a particular QI, analytically Auger et al. (2009b); Brockhoff et al. (2012); Lizárraga-Lizárraga et al. (2008c) or empirically Ishibuchi et al. (2018b, 2015, 2014), the effectiveness investigation of a group of QIs in benchmark functions Knowles et al. (2006); Okabe et al. (2003) or in real-world applications Wang et al. (2016), and also the correlation analysis of different QIs Jiang et al. (2014); Liefooghe and Derbel (2016); Ravber et al. (2017).

5. Review of certain aspects of QIs. Work in this category focuses on overview from specific perspectives, such as a review of QIs for exact multiobjective optimisers Faulkenberg and Wiecek (2010), a critical review of several well-established QIs Knowles (2002); Knowles and Corne (2002); Okabe et al. (2003), a statistical review of the use frequency of QIs in the literature Riquelme et al. (2015); Wang et al. (2016), and an instructional review on how to design QIs Hansen and Jaszkiewicz (1998); Knowles et al. (2006) or on how to choose suitable QIs for particular problems Li et al. (2018a); Wang et al. (2016).

From the above, however, it can be seen that there is no comprehensive review of the quality evaluation studies — existing works focus on one or several particular facets of quality evaluation, e.g., a summary of some QIs Faulkenberg and Wiecek (2010); Riquelme et al. (2015), a criticism of misleading results obtained by QIs Knowles and Corne (2002); Okabe et al. (2003), a discussion of some issues in developing QIs Zitzler et al. (2008, 2003), a general guide of designing QIs Hansen and Jaszkiewicz (1998); Knowles et al. (2006), and a practical guide of selecting suitable QIs in a particular class of problems Li et al. (2018a); Wang et al. (2016). This contrasts with a range of comprehensive reviews in other topics in multiobjective optimisation, such as solving multiobjective optimisation by evolutionary algorithms Coello (2000); Zhou et al. (2011), decomposition-based multi-objective evolutionary algorithms Trivedi et al. (2017), evolutionary many-objective optimisation Li et al. (2015), multiobjective approaches to data mining Mukhopadhyay et al. (2014a,b), multiobjective approaches to finance and economics applications Ponsich et al. (2013), multiple criteria decision making Wallenius et al. (2008), and dynamic multiobjective optimisation benchmark Helbig and Engelbrecht (2014).

This paper attempts to fill this gap by covering all important facets of quality evaluation. We systematically review 100 QIs, analyse the strengths and weaknesses of representative QIs, discuss main issues in quality evaluation, give detailed recommendations in designing, selecting and using QIs, and suggest several future research directions in quality evaluation.

Note that this review focuses on the evaluation of solution sets rather than on multiobjective optimisers. This thus does not involve the comparison of running time of optimisers, statistical results

of stochastic optimiser in multiple runs, and other generation/iteration-wise indicators. Readers who are interested in those respects can refer to Datta and Figueira (2012); Fonseca and Fleming (1996); Knowles et al. (2006); Li and Yao (2019); Van Veldhuizen (1999); Zitzler et al. (2008).

The rest of the paper is outlined as follows. Section 2 introduces the background of quality evaluation in multiobjective optimisation. Section 3 systematically reviews 100 QIs in the literature and details several representative QIs. Section 4 highlights several important issues in quality evaluation. Future research directions are suggested in Section 5, and finally Section 6 concludes the paper.

2 Background

In this section, we first give basic concepts in multiobjective optimisation and common comparison relations between solution sets. We then describe general aspects of solution set quality, which is followed by the introduction of quality evaluation of solution sets.

2.1 Terminology

Without loss of generality, we consider a minimisation MOP with n decision variables and m objective functions $f : X \to Z$, $X \subset \mathbb{R}^n$, $Z \subset \mathbb{R}^m$. The objective functions map a point $\mathbf{x} \in X$ in the decision space to an objective vector $f(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_m(\mathbf{x}))$ in the objective space. This paper focuses on quality evaluation of objective vector sets, and the comparison relation is defined on the basis of objective vectors. In addition, for simplicity we refer to an objective vector as a *solution* (despite it being originally termed in X) and the outcome of a multiobjective optimiser as a *solution set*. We mainly consider the most general case, namely, no additional knowledge about the MOP available, e.g., the preference information of the decision maker (DM) unknown *a priori*.

Considering two solutions $\mathbf{x}, \mathbf{y} \in Z$, solution $\mathbf{x}$ is said to *weakly (Pareto) dominate* $\mathbf{y}$ (denoted as $\mathbf{x} \preceq \mathbf{y}$) if $\mathbf{x}_i \leq \mathbf{y}_i$ for $1 \leq i \leq m$. If there exists at least one objective j on which $\mathbf{x}_j < \mathbf{y}_j$, we say that $\mathbf{x}$ *dominates* $\mathbf{y}$ (denoted as $\mathbf{x} \prec \mathbf{y}$). A solution $\mathbf{x} \in Z$ is called *Pareto optimal* (or *efficient*) if there is no $\mathbf{y} \in Z$ that dominates $\mathbf{x}$. The set of all Pareto optimal solutions of an MOP is called its *Pareto front* (or *Pareto optimal frontier*). In addition, there exist relations stricter than Pareto dominance, e.g., *strict Pareto dominance*. Solution $\mathbf{x}$ is said to *strictly (Pareto) dominate* $\mathbf{y}$ (denoted as $\mathbf{x} \prec\prec \mathbf{y}$) if $\mathbf{x}_i < \mathbf{y}_i$ for $1 \leq i \leq m$.

The relations between solutions can be readily extended to between solution sets. Let $\mathbf{A}$ and $\mathbf{B}$ be two nondominated solution sets (i.e., in each solution set, all solutions are nondominated with each other).

Definition 1 (Weak Dominance Zitzler et al. (2003))**.** *Set* $\mathbf{A}$ *is said to weakly dominate* $\mathbf{B}$ *(denoted as* $\mathbf{A} \preceq \mathbf{B}$*) if every solution* $\mathbf{b} \in \mathbf{B}$ *is weakly dominated by at least one solution* $\mathbf{a} \in \mathbf{A}$*.*

Definition 2 (Dominance Zitzler et al. (2003))**.** *Set* $\mathbf{A}$ *is said to dominate* $\mathbf{B}$ *(denoted as* $\mathbf{A} \prec \mathbf{B}$*) if every solution* $\mathbf{b} \in \mathbf{B}$ *is dominated by at least one solution* $\mathbf{a} \in \mathbf{A}$*.*

This relation is also termed *complete outperformance* ($A \; \mathcal{O}_{\mathcal{C}} \; B$) in Hansen and Jaszkiewicz (1998).

Definition 3 (Strict Dominance Zitzler et al. (2003))**.** *Set* $\mathbf{A}$ *is said to strictly dominate* $\mathbf{B}$ *(denoted as* $\mathbf{A} \prec\prec \mathbf{B}$*) if every solution* $\mathbf{b} \in \mathbf{B}$ *is strictly dominated by at least one solution* $\mathbf{a} \in \mathbf{A}$*.*

On top of these three relations, there are relations concerning exclusively the comparison between solution sets.

Definition 4 (Strong Outperformance Hansen and Jaszkiewicz (1998))**.** *Set* $\mathbf{A}$ *is said to strongly outperform* $\mathbf{B}$ *(denoted as* $\mathbf{A} \; \mathcal{O}_{\mathcal{S}} \; \mathbf{B}$*) if* $\mathbf{A} \preceq \mathbf{B}$ *and also there exist at least one pair of solutions* $\mathbf{a} \in \mathbf{A}$ *and* $\mathbf{b} \in \mathbf{B}$ *such that* $\mathbf{a} \prec \mathbf{b}$*.*

Table 1: Common relations on two solution sets $\mathbf{A}$ and $\mathbf{B}$, where "dominance" terms are from Zitzler et al. (2003) and "outperformance" terms are from Hansen and Jaszkiewicz (1998)

Relation	Symbol	Definition
weakly dominate	$\mathbf{A} \preceq \mathbf{B}$	$\forall \mathbf{b} \in \mathbf{B}, \exists \mathbf{a} \in \mathbf{A}, \mathbf{a} \preceq \mathbf{b}$
better (or weakly outperform)	$\mathbf{A} \lhd \mathbf{B}$ (or $\mathbf{A}\ \mathcal{O}_{\mathcal{W}}\ \mathbf{B}$)	$\mathbf{A} \preceq \mathbf{B} \wedge \mathbf{A} \neq \mathbf{B}$
strongly outperform	$\mathbf{A}\ \mathcal{O}_{\mathcal{S}}\ \mathbf{B}$	$\mathbf{A} \preceq \mathbf{B} \wedge \exists \mathbf{b} \in \mathbf{B}, \exists \mathbf{a} \in \mathbf{A}, \mathbf{a} \prec \mathbf{b}$
dominate (or completely outperform)	$\mathbf{A} \prec \mathbf{B}$ (or $\mathbf{A}\ \mathcal{O}_{\mathcal{C}}\ \mathbf{B}$)	$\forall \mathbf{b} \in \mathbf{B}, \exists \mathbf{a} \in \mathbf{A}, \mathbf{a} \prec \mathbf{b}$
strictly dominate	$\mathbf{A} \prec\prec \mathbf{B}$	$\forall \mathbf{b} \in \mathbf{B}, \exists \mathbf{a} \in \mathbf{A}, \mathbf{a} \prec\prec \mathbf{b}$

Definition 5 (Better Zitzler et al. (2003))**.** *Set* $\mathbf{A}$ *is said to be better than* $\mathbf{B}$ *(denoted as* $\mathbf{A} \lhd \mathbf{B}$*) if* $\mathbf{A} \preceq \mathbf{B}$ *and also there exists at least one solution* $\mathbf{a} \in \mathbf{A}$ *that is not weakly dominated by any solution in* $\mathbf{B}$*.*

This relation is also termed *weak outperformance* ($\mathbf{A}\ \mathcal{O}_{\mathcal{W}}\ \mathbf{B}$) in Hansen and Jaszkiewicz (1998). It represents the most general and weakest form of superiority between two solution sets, namely, $\mathbf{A} \preceq \mathbf{B}$ but $\mathbf{B} \npreceq \mathbf{A}$. In other words, $\mathbf{A}$ is at least as good as $\mathbf{B}$, while $\mathbf{B}$ is not as good as $\mathbf{A}$.

There exists an ordering amongst the above relations as $\mathbf{A} \prec\prec \mathbf{B} \Rightarrow \mathbf{A} \prec \mathbf{B} \Rightarrow \mathbf{A}\ \mathcal{O}_{\mathcal{S}}\ \mathbf{B} \Rightarrow \mathbf{A} \lhd \mathbf{B} \Rightarrow \mathbf{A} \preceq \mathbf{B}$. The difference between $\mathcal{O}_{\mathcal{S}}$ and $\lhd$ is that $\mathbf{A} \lhd \mathbf{B}$ includes the situation that $\mathbf{B}$ is a proper subset of $\mathbf{A}$ ($\mathbf{B} \subset \mathbf{A}$), but $\mathbf{A}\ \mathcal{O}_{\mathcal{S}}\ \mathbf{B}$ does not. Table 1 summarises these relations.

2.2 Solution Set Quality

As mentioned before, the *better* relation $\lhd$ is the most general assumption of the DM's preferences to compare solution sets. It meets any preference potentially articulated by the DM (when the concept of optimum is solely based on Pareto dominance, other than on other criteria, e.g., robustness). However, the *better* relation may leave many solution sets incomparable as there likely exist some solutions being nondominated with each other in the sets. This naturally requires stronger assumptions about the DM's preference in distinguishing between solution sets. Yet, stronger assumptions cannot guarantee that the favoured set (under the assumptions) is certainly preferred by the DM. This is not surprising as different DMs indeed may prefer different trade-offs amongst objectives. Nevertheless, when the DM's preferences are unknown *a priori*, a solution set having a sufficient number of nondominated solutions, good closeness to the Pareto front, good spread over the Pareto front, and good uniformity amongst solutions is often preferable, since it can well represent the Pareto front and thus has a greater probability of being preferred by the DM. Moreover, a good representation of the Pareto front can reveal the problem's properties, such as the shape, actual dimensionality, scale, knee point, and relationship between objectives, such that the user can understand better the problem she/he is dealing with. In general, the quality of a solution set can be interpreted as how well it represents the Pareto front, and can be broken down into four aspects: *convergence*, *spread*, *uniformity*, and *cardinality*.

Convergence of a solution set refers to the closeness of the set to the Pareto front. Spread of a solution set considers the region of the set covering. It involves both outer portion and inner portion of the set. This is not like the quality *extensity* which merely takes the boundaries of the set into account. Note that the spread of a solution set, in the presence of the problem's Pareto front, is also known as the *coverage* of the set Sayın (2000). Uniformity of a set refers to how even the solution distribution is in the set; an equidistant spacing amongst solutions is desirable. Spread and uniformity are closely related, and they collectively are known as the *diversity* of a set. Cardinality of a solution set refers to the number of solutions in the set. In general, we desire sufficient solutions to clearly describe the set, but not too many that may overwhelm the DM with choices. Nevertheless, it is believed that a set having a larger number of solutions is preferred if two sets are generated with the same amount of computational resources.

2.3 Quality Comparison

One straightforward way to compare the quality of solution sets is to visualise the sets and judge intuitively the superiority of one set to another. Such visual comparison is one of the most frequently used methods and it is well suited to bi- or tri-objective MOPs. When the number of objectives is larger than three, where the direct observation of solution sets is unavailable (by *scatter plot*), people may resort to the tools from the data analysis field, e.g., *parallel coordinates* Inselberg and Dimsdale (1991), *spider-web charts* Kasanen et al. (1991), and *scatter plot matrix* Tukey and Tukey (1981) (see Miettinen (2014) for a summary), or adopt visualisation techniques developed specifically for multi-objective optimisation, e.g., barycentric coordinates-based *RadViz* Walker et al. (2013), *prosection* method Tusar and Filipic (2015), and *polar-metric* He and Yen (2016). However, these visualisation methods may not be able to clearly reflect all the aspects of solution set quality; for example, commonly-used parallel coordinates only partially reflect the convergence, spread and uniformity Li et al. (2017). In addition, visual comparison cannot quantify the difference between solution sets, and thus cannot be used to guide the optimisation.

Quality indicators (QIs) overcome the issues of visual comparison by mapping a solution set into a real number, thereby providing quantitative differences between solution sets. QIs are capable of delivering precise statements of solution set quality, for example, in which quality aspect one set is better than another and how much better is one set than another in certain aspects. In principle, any function of mapping a set of vectors into a scalar value can be seen as a potential quality indicator, but in general it may need to reflect one or several aspects of set quality: convergence, spread, uniformity, and cardinality. Note that when comparing solution sets generated by exact methods, the convergence evaluation of solution sets is excluded since the generated solution set is a subset of the problem's Pareto front.

3 Overview of Quality Indicators in the Literature

This section reviews QIs on the basis of what quality aspects they are mainly capturing. In general, QIs can be divided into six categories — 1) QIs for convergence, 2) QIs for spread, 3) QIs for uniformity, 4) QIs for cardinality, 5) QIs for both spread and uniformity, and 6) QIs for combined quality of the four quality aspects. In each category, we also detail one or several example indicators. These QIs are commonly used in the literature and/or are representative in their category. Table 2 summarises all 100 QIs in the literature. Note that it does not include measures which are a combination of several QIs, such as those in Yen and He (2014).

3.1 QIs for Convergence

As the most important aspect of a solution set's quality, convergence has received a lot of attention in set evaluation. There exist two classes of convergence QIs in the literature. One is to consider the Pareto dominance relation between solutions or sets (items 1–9 in Table 2); the other is to consider the distance of a solution set to the Pareto front or one/several points derived from the Pareto front (items 10–22).

3.1.1 Dominance-based QIs

A type of frequently-used dominance-based QIs is to consider the dominance relation between solutions of two sets, such as the $\mathcal{C}$ indicator Zitzler and Thiele (1998), $\hat{\mathcal{C}}$ indicator Fieldsend et al. (2003), σ-, τ- and κ metrics Datta and Figueira (2012), and *contribution indicator* Meunier et al. (2000). Other QIs concerning solutions' dominance include *wave metric* Van Veldhuizen (1999), *purity* Bandyopadhyay et al. (2004), *Pareto dominance indicator* Goh and Tan (2009), and *dominance-based quality* Bui et al. (2009). The wave metric crunches the number of the nondominated fronts

in a solution set. The purity indicator counts nondominated solutions of the considered set over the combined collection of all the candidate sets. The Pareto dominance indicator measures the ratio of the combined set's nondominated solutions that are contributed by a particular set. The dominance-based quality considers the dominance relation between a solution and its neighbours in the set.

The above QIs are all based on the dominance relation between solutions. This contrasts with *dominance ranking* which is based on the dominance relation between sets. The dominance ranking indicator considers the combined collection of all the sets and assigns each set a rank on the basis of a dominance criterion, for example, *dominance count* Fonseca and Fleming (1995) or *nondominated sorting* Goldberg (1989).

However, all dominance-based QIs have some weaknesses. They provide little information about what extent one set outperforms another. More importantly, they may leave solution sets incomparable if all solutions of the sets are nondominated to each other, which may happen frequently in many-objective optimisation Ishibuchi et al. (2008); Li et al. (2015b); Purshouse and Fleming (2007). In addition, it is worth noting that some dominance-based QIs may partially imply the cardinality of a solution set since a bigger-size set may result in more solutions nondominated, such as the $\mathcal{C}$ Zitzler and Thiele (1998), contribution indicator Meunier et al. (2000), and Pareto dominance indicator Goh and Tan (2009).

Table 2: Quality indicators and their properties. "+" generally means that the indicator can well reflect the specified quality of solution sets. "−" for convergence means that the indicator can reflect the convergence of a set to some extent; e.g., indicators only considering the dominance relation as convergence measure. "−" for spread means that the indicator can only reflect the extensity of a set. "−" for uniformity means that the indicator can reflect the uniformity of a set to some extent; i.e., a disturbance to an equally-spaced set may not certainly lead to a worse evaluation result. "−" for cardinality means that adding a nondominated solution into a set is not surely but likely to lead to a better evaluation result and also it never leads to a worse evaluation result.

No.	Quality indicator	Convergence	Spread	Uniformity	Cardinality
1	$\mathcal{C}$ Zitzler and Thiele (1998)	−			−
2	$\hat{\mathcal{C}}$ Fieldsend et al. (2003)	−			−
3	Contribution indicator (CI) Meunier et al. (2000)	−			−
4	Dominance-based quality Bui et al. (2009)	−			
5	Dominance ranking Knowles et al. (2006)	−			
6	Pareto dominance indicator Goh and Tan (2009)	−			−
7	Purity Bandyopadhyay et al. (2004)	−			−
8	Wave metric Van Veldhuizen (1999)	−			
9	σ-, τ- and κ metrics Datta and Figueira (2012)	−			−
10	DistZ Viana and de Sousa (2000)	−			
11	Distance to a knee point Emmerich et al. (2007)	−	−		
12	Distance to the ideal point (ED) Zeleny (1973)	−			
13	Tchebycheff distance to the knee point Jaimes and Coello (2009)	−	−		
14	Seven point average distance Schott (1995)	+			
15	Convergence index Nicolini (2004)	+			
16	Convergence metric (CM) Deb and Jain (2002)	+			
17	Generational distance (GD) Van Veldhuizen and Lamont (1998)	+			
18	GD_p Schutze et al. (2012)	+			
19	GD^+ Ishibuchi et al. (2015)	+			
20	$\mathcal{M}_1^*$ Zitzler et al. (2000)	+			
21	Maximum Pareto front error Van Veldhuizen (1999)	+			
22	Mean absolute error Kaji and Kita (2007)	+			
23	Area & length De et al. (1992)		+		
24	Coverage error ϵ Sayın (2000)		+		−
25	Extension Meng et al. (2005)		−		
26	$\mathcal{M}_3^*$ (Maximum spread or MS) Zitzler et al. (2000)		−		
27	Modified MS Adra and Fleming (2011)		−		
28	MS' Goh and Tan (2007)		−		
29	Outer diameter Zitzler et al. (2008)		−		
30	Overall Pareto spread Wu and Azarm (2001)		−		
31	PD Wang et al. (2017)		+		
32	Spread assessment Li and Zheng (2009)		−		

Table 2: Quality indicators and their properties. "+" generally means that the indicator can well reflect the specified quality of solution sets. "−" for convergence means that the indicator can reflect the convergence of a set to some extent; e.g., indicators only considering the dominance relation as convergence measure. "−" for spread means that the indicator can only reflect the extensity of a set. "−" for uniformity means that the indicator can reflect the uniformity of a set to some extent; i.e., a disturbance to an equally-spaced set may not certainly lead to a worse evaluation result. "−" for cardinality means that adding a nondominated solution into a set is not surely but likely to lead to a better evaluation result and also it never leads to a worse evaluation result.

No.	Quality indicator	Convergence	Spread	Uniformity	Cardinality
33	Spread measure Ishibuchi and Shibata (2004)		−		
34	Cluster Wu and Azarm (2001)			+	
35	Deviation measure Δ Deb et al. (2000)			+	
36	Evenness Messac and Mattson (2004)			+	
37	Hole relative size Collette and Siarry (2005)			+	
38	Minimal spacing Bandyopadhyay et al. (2004)			+	
39	Spacing (SP) Schott (1995)			+	
40	Spacing measure Collette and Siarry (2005)			−	
41	Uniformity Meng et al. (2005)			+	
42	Uniformity assessment Li et al. (2008)			+	
43	Uniformity distribution Tan et al. (2002)			+	
44	Uniformity level δ Sayın (2000)			+	
45	Cardinality D Sayın (2000)				+
46	Number of unique nondominated solutions Berry and Vamplew (2005)				+
47	Overall nondominated vector generation (ONVG) Van Veldhuizen (1999)				+
48	Ratio of nondominated individuals Tan et al. (2002)				+
49	$C1$ Hansen and Jaszkiewicz (1998)				−
50	$C2$ Hansen and Jaszkiewicz (1998)				−
51	Error ratio Van Veldhuizen (1999)				+
52	ONVG ratio Van Veldhuizen (1999)				+
53	Proportion of Pareto-optimal objective vectors found Ulungu et al. (1999)				−
54	Success counting Sierra and Coello (2005)				−
55	Cluster-based diversity metric Li et al. (2005)		+	−	
56	Coverage over Pareto front (CPF) Tian et al. (2019)		+	+	
57	HV_d Jiang et al. (2016)		+	+	+
58	Relative entropy Meunier et al. (2000)		+	−	−
59	Extended spread Zhou et al. (2006)		−	+	
60	Sparsity index Nicolini (2004)		−	−	
61	Δ Deb et al. (2002)		+	+	
62	Δ_{Line} Ibrahim et al. (2017)		+	−	
63	Chi-square-like deviation Srinivas and Deb (1994)		+	−	
64	Cover rate Hiroyasu et al. (2000)		+	−	
65	Diversity comparison indicator (DCI) Li et al. (2014a)		+	−	−
66	Diversity metric (DM) Deb and Jain (2002)		+	−	−
67	DIR Cai et al. (2018)		+	−	−
68	Entropy Farhang-Mehr and Azarm (2003a)		+	−	−
69	$\mathcal{M}_2^*$ Zitzler et al. (2000)		+	−	
70	M-DI Asafuddoula et al. (2015)		+	−	−
71	Number of distinct choices Wu and Azarm (2001)		+	−	−
72	Sigma diversity metric Mostaghim and Teich (2005)		+	−	−
73	Sparsity Deb et al. (2005)		+	−	
74	U-measure Leung and Wang (2003)		+	+	
75	Averaged Hausdorff distance Δ_p Schutze et al. (2012)	+	+	−	−
76	Dist1 ($D1$) Czyzak and Jaszkiewicz (1998)	+	+	−	−
77	Dist2 ($D2$) Czyzak and Jaszkiewicz (1998)	+	+	−	−
78	Degree of approximation (DOA) Dilettoso et al. (2017)	+	+	−	−
79	Delineation of Pareto optimal front Eskandari et al. (2007)	+	+	−	−
80	Dominance move (DoM) Li et al. (2019)	+	+	−	−
81	Epsilon indicator (ϵ-indicator) Zitzler et al. (2003)	+	+	−	−
82	ϵ performance Kollat and Reed (2005)	+	+	−	−
83	Front to set distance Bosman and Thierens (2005)	+	+	−	−
84	G-Metric Lizárraga-Lizárraga et al. (2008a)	−	−	+	
85	Inverted generational distance (IGD) Coello and Sierra (2004)	+	+	−	−
86	IGD_p Schutze et al. (2012)	+	+	−	−
87	IGD^+ Ishibuchi et al. (2015)	+	+	−	−
88	IGD-NS Tian et al. (2016)	+	+	−	−
89	I_{SDE} Li et al. (2016, 2014b)	+	+	−	
90	ObjIGD Ibrahim et al. (2017)	+	+	−	−
91	Performance comparison indicator (PCI) Li et al. (2015a)	+	+	−	−

Table 2: Quality indicators and their properties. "+" generally means that the indicator can well reflect the specified quality of solution sets. "−" for convergence means that the indicator can reflect the convergence of a set to some extent; e.g., indicators only considering the dominance relation as convergence measure. "−" for spread means that the indicator can only reflect the extensity of a set. "−" for uniformity means that the indicator can reflect the uniformity of a set to some extent; i.e., a disturbance to an equally-spaced set may not certainly lead to a worse evaluation result. "−" for cardinality means that adding a nondominated solution into a set is not surely but likely to lead to a better evaluation result and also it never leads to a worse evaluation result.

No.	Quality indicator	Convergence	Spread	Uniformity	Cardinality
92	Completeness indicator Lotov et al. (2013, 2002)	+	+	−	−
93	Coverage difference Zitzler (1999)	+	+	−	+
94	Hyperarea difference Wu and Azarm (2001)	+	+	−	+
95	Hyperarea ratio Van Veldhuizen (1999)	+	+	−	+
96	Hypervolume (HV) Zitzler and Thiele (1998)	+	+	−	+
97	Integrated preference functional (IPF) Carlyle et al. (2003)	+	+	−	−
98	IPF with Tchebycheff function Bozkurt et al. (2010)	+	+	−	−
99	$R1$, $R2$, $R3$ Hansen and Jaszkiewicz (1998)	+	+	−	−
100	Volume measure Fieldsend et al. (2003)	+	+	−	+

3.1.2 Distance-based QIs

The majority of QIs for convergence falls into this class (items 10–22 in Table 2). They can further be classified into two groups. One is to measure the distance of the considered solution set to one or several particular points derived from the Pareto front (items 10–14), such as the ideal point Zeleny (1973), knee point(s) Emmerich et al. (2007); Jaimes and Coello (2009), the Zeleny point Viana and de Sousa (2000) and the seven particular points Schott (1995). The ideal point is the point constructed by the best value on each objective of the Pareto front. A knee point is the point on the Pareto front having the maximum reflex angle computed from its neighbours. The Zeleny point is the point obtained by minimising each objective separately. The seven points defined in Schott (1995) are seven particular points derived from the ideal point and the extreme points of the Pareto front for bi-objective problems.

The other group is to measure the distance to a reference set[2] that well represents the Pareto front (items 15–22). In this group, the indicator *GD* Van Veldhuizen and Lamont (1998) is most frequently used. GD first calculates the Euclidean distance for each solution in the solution set to the closest point in the reference set, and then takes the quadratic mean over all of these distances. Other QIs in this group can be seen as variants of GD Kaji and Kita (2007); Nicolini (2004), for example, taking the arithmetic mean of the distances Deb and Jain (2002); Zitzler et al. (2000) and the power mean Schutze et al. (2012), considering the Tchebycheff distance Van Veldhuizen (1999) and introducing the dominance relation between solutions and points in the reference set Ishibuchi et al. (2015).

In the next two sections, we will introduce two representative convergence QIs in detail, one based on dominance and the other based on distance.

3.1.3 $\mathcal{C}$ Indicator

The $\mathcal{C}$ indicator Zitzler and Thiele (1998) is arguably the best-known dominance-based QI. It considers the dominance relation between two solution sets and measures their relative quality on convergence and cardinality. Given two sets $\mathbf{A}$ and $\mathbf{B}$, $\mathcal{C}(\mathbf{A},\mathbf{B})$ gauges the proportion of solutions of $\mathbf{B}$ that are weakly dominated by at least one solution of $\mathbf{A}$. Formally,

$$\mathcal{C}(\mathbf{A},\mathbf{B}) = \frac{|\mathbf{b} \in \mathbf{B} \mid \exists \mathbf{a} \in \mathbf{A} : \mathbf{a} \preceq \mathbf{b}|}{\mathbf{B}} \tag{1}$$

[2]For benchmarking functions, the reference set is typically constructed by a set of densely and uniformly distributed points over the Pareto front; for real world problems whose Pareto front is unknown, the reference set often consists of the nondominated solutions of the collections of all solutions produced during the search.

$\mathcal{C}(\mathbf{A}, \mathbf{B})$ maps the pair of $\mathbf{A}$ and $\mathbf{B}$ to the interval $[0, 1]$. $\mathcal{C}(\mathbf{A}, \mathbf{B}) = 0$ means that none of the solutions in $\mathbf{B}$ is weakly dominated by any member of $\mathbf{A}$, and $\mathcal{C}(\mathbf{A}, \mathbf{B}) = 1$ means $\mathbf{A} \preceq \mathbf{B}$.

Note that in quality comparison of two sets, both $\mathcal{C}(\mathbf{A}, \mathbf{B})$ and $\mathcal{C}(\mathbf{B}, \mathbf{A})$ need to be considered since $\mathcal{C}(\mathbf{A}, \mathbf{B}) \neq 1 - \mathcal{C}(\mathbf{B}, \mathbf{A})$ unless the two sets have no solutions being nondominated to each other. Another issue of the indicator is that due to the use of the "$\preceq$" relation in the evaluation, $\mathcal{C}(\mathbf{A}, \mathbf{A}) \neq 0$ and two nondominated sets can take any value in $[0, 1]$ (depending on how many repetitive solutions they have in common). Although a replacement by the "$\prec$" relation can overcome the above problems Fieldsend et al. (2003), the resulting version does not satisfy the triangle inequality. In addition, the $\mathcal{C}$ indicator has also been extended by taking into account the dominated solutions Meunier et al. (2000) or by introducing other dominance relations and comparing any number of sets Datta and Figueira (2012).

3.1.4 Generational Distance (GD)

As aforementioned, the GD indicator Van Veldhuizen and Lamont (1998) measures the quadratic mean of the Euclidean distances of solutions in the given set to the closest point on the Pareto front. Formally, given a solution set $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N\}$,

$$GD(\mathbf{A}) = \frac{1}{N} \left(\sum_{i=1}^{N} (d_2(\mathbf{a}_i, \mathrm{PF}))^2 \right)^{1/2} \tag{2}$$

where $d_2(\mathbf{a}_i, \mathrm{PF})$ means the L^2 norm distance (Euclidean distance) of solution $\mathbf{a}_i$ to the Pareto front. Usually, a reference set $\mathbf{R}$ that well represents the Pareto front is used in practice, thus

$$d_2(\mathbf{a}_i, \mathrm{PF}) = \min_{\mathbf{r} \in \mathbf{R}} d_2(\mathbf{a}_i, \mathbf{r})$$

where $d_2(\mathbf{a}_i, \mathbf{r})$ denotes the Euclidean distance between $\mathbf{a}_i$ and $\mathbf{r}$. Of course, GD may not necessarily require a reference set representing the Pareto front if the geometric property of the front is known.

The GD value is to be minimised; a result of zero indicates that the set is on the Pareto front/reference set. As designed for generational evaluation, GD is often used to measure evolutionary progress of the solution set towards the Pareto front. However, since GD considers the *quadratic mean*, it is rather sensitive to outliers and returns a solution set having outliers a poor score no matter how the rest performs. In addition, GD can be affected by the size of the solution set Schutze et al. (2012). When $N \to \infty$, GD $\to 0$ even if the set is far away from the Pareto front. Therefore, GD is reliably usable only when the sets under consideration have the same/or very similar size. Fortunately, this issue can be fixed if replacing the *quadratic mean* with the *arithmetic mean* in Equation (2). In fact, in some recent studies (e.g., Schutze et al. (2012) and Ishibuchi et al. (2015)), a general form of the GD indicator with the exponent "p" and "$1/p$" instead of "2" and "1/2" was adopted. Setting $p = 1$ has now been commonly accepted and used in line with its inverted version, IGD (cf. Equation (6)) which measures the arithmetic mean of the distances from points of the Pareto front to the closest solution in the considered set.

3.2 QIs for Spread

Spread quality is concerned with the area of a solution set covering. A set with good spread should contain solutions from every portion of the Pareto front without missing out any region. However, most spread QIs only measure the extent of a solution set. Table 2 lists 11 spread QIs in the literature (items 23–33). These QIs typically consider the range formed by the extreme solutions of the set, such as *maximum spread* (MS) Zitzler et al. (2000) and its variants Adra and Fleming (2011); Goh and Tan (2007); Ishibuchi and Shibata (2004); Meng et al. (2005); Wu and Azarm (2001); Zitzler

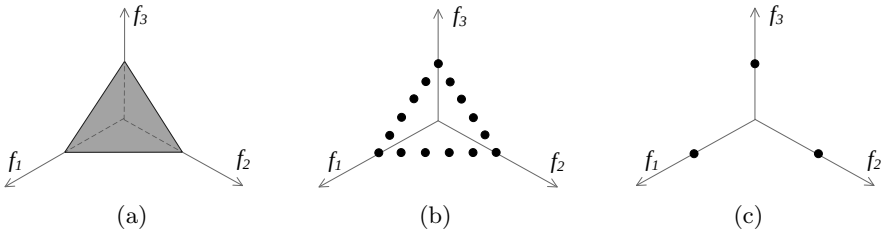

Figure 2: A tri-objective example of boundary solutions and extreme solutions of a Pareto front. (a) Pareto front, (b) boundary solutions, and (c) extreme solutions.

et al. (2008), or consider the range enclosed by the boundary solutions of the set, such as *spread assessment* Li and Zheng (2009). We borrow a figure in Li et al. (2014) to illustrate the extreme solutions and the boundary solutions. As seen in Figure 2, QIs that only take these solutions into account could miss the inner regions of the Pareto front.

Fortunately, there do exist some QIs designed for the whole coverage of the solution set. For example, *area & length* De et al. (1992) measures the area and length of the *supported points* of a solution set. *Coverage error* ϵ Sayın (2000) calculates the maximum dissimilarity of a solution set over the Pareto front. *PD* Wang et al. (2017) sums up the dissimilarity of each solution to the remaining solutions of a solution set.

3.2.1 Maximum Spread (MS)

MS (or $\mathcal{M}_3^*$) Zitzler et al. (2000) is a widely used spread indicator, and it measures the range of a solution set by considering the maximum extent on each objective. MS is defined as

$$MS(\mathbf{A}) = \sqrt{\sum_{j=1}^{m} \max_{\mathbf{a},\mathbf{a}' \in \mathbf{A}} (\mathbf{a}_j - \mathbf{a}'_j)^2} \tag{3}$$

where m denotes the number of objectives. MS is to be maximised; the higher the value, the better extensity to be claimed. In the case of bi-objective scenarios, the MS value of a nondominated solution set is the Euclidean distance of its two extreme solutions.

However, as mentioned previously, MS which only considers the extreme solutions of the set fails to reflect the spread quality. In addition, as it does not touch on the convergence of the set, the solutions that are far away from the Pareto front usually contribute a lot to the MS value. This easily induces misleading evaluation. For example, a solution set that concentrates on a tiny portion of the Pareto front but has one outlier far away from the front would be assigned a good MS value. To deal with this, the range of the Pareto front is brought in as a reference in the evaluation, e.g, MS' Goh and Tan (2007) and modified MS Adra and Fleming (2011).

3.3 QIs for Uniformity

Quality indicators for uniformity measure how uniformly a set's solutions are distributed. Since the quality of a solution set can be seen as its ability of representing the Pareto front, a uniformly distributed solution set, which provides a better Pareto front representation than a non-uniformly one, can be considered to possess better quality. A desirable uniformity QI should rank highest to a set consisting of solutions that are spaced completely equally to each other, and a little disturbance to this set should lead to a worse evaluation result. Items 34–44 in Table 2 correspond to uniformity QIs.

The uniformity quality of a solution set can typically be evaluated by measuring the variation of the distance between the solutions. Many QIs in this class are designed along these lines, such as

spacing (SP) Schott (1995), *deviation measure* Δ Deb et al. (2000), *uniformity distribution* Tan et al. (2002), *minimal spacing* Bandyopadhyay et al. (2004), *spacing measure* Collette and Siarry (2005) and *uniformity* Meng et al. (2005). Other QIs for uniformity include considering the minimum/maximum distance between solutions Collette and Siarry (2005); Messac and Mattson (2004); Sayın (2000), and constructing clusters Wu and Azarm (2001) or a minimum spanning tree Li et al. (2008) (of all the solutions of the set under consideration) to evaluate the uniformity quality.

It is worth mentioning that having equidistant solutions for a set does not guarantee having good diversity. As such, QIs for uniformity should always be used in conjunction with a spread QI.

3.3.1 Spacing (SP)

As the most popular uniformity indicator, SP Schott (1995) gauges the variation of the distance between solutions in a set. Specifically, given a solution set $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N\}$,

$$SP(\mathbf{A}) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\overline{d} - d_1(\mathbf{a}_i, \mathbf{A}/\mathbf{a}_i))^2} \tag{4}$$

where $\overline{d}$ is the mean of all $d_1(\mathbf{a}_1, \mathbf{A}/\mathbf{a}_1), d_1(\mathbf{a}_2, \mathbf{A}/\mathbf{a}_2), ..., d_1(\mathbf{a}_N, \mathbf{A}/\mathbf{a}_N)$, and $d_1(\mathbf{a}_i, \mathbf{A}/\mathbf{a}_i)$ means the L^1 norm distance (Manhattan distance) of $\mathbf{a}_i$ to the set $\mathbf{A}/\mathbf{a}_i$, namely,

$$d_1(\mathbf{a}_i, \mathbf{A}/\mathbf{a}_i) = \min_{\mathbf{a}\in\mathbf{A}/\mathbf{a}_i}\sum_{j=1}^{m}|\mathbf{a}_{ij} - \mathbf{a}_j|$$

where m denotes the number of objectives and $\mathbf{a}_{ij}$ is the jth objective of solution $\mathbf{a}_i$. SP is to be minimised; the lower the value, the better the uniformity. A SP value of zero indicates all members of the solution set are spaced equidistantly on the basis of Manhattan distance. Note that SP only gauges the distribution in the "neighbourhood" of solutions. Even if working with MS, SP cannot cover the diversity quality of the set, though these two indicators were often used together to serve this purpose in the literature. Taking Figure 2 as an example, the solution sets in both Figures 2(b) and (c) will take full scores by SP and MS; however they are located only in boundaries and extremes of the Pareto front, respectively.

3.4 QIs for Cardinality

QIs for cardinality can boil down to a simple idea — counting the number of nondominated solutions. A desirable (necessary) property that cardinality-based QIs possess is that adding a different nondominated solution to the set under consideration should improve (does not degrade) the evaluation result. This is in line with the concept of *(weak) monotony* in Knowles and Corne (2002). Table 2 lists 10 cardinality QIs (items 45–54).

Depending on involvement of the Pareto optimal solutions, cardinality-based QIs can be grouped into two classes (items 45–48 and items 49–54). One is to directly consider the nondominated solutions in a set, e.g., the indicators *cardinality D* Sayın (2000), *number of unique nondominated solutions* Berry and Vamplew (2005), *overall nondominated vector generation (ONVG)* Van Veldhuizen (1999) and *ratio of nondominated individuals* Tan et al. (2002). The other class makes a comparison between nondominated solutions in the set and the Pareto optimal solutions of the problem. QIs in this class typically return a ratio of the nondominated solutions that belong to the Pareto optimal set to the size of the optimal set (e.g., $C1$ Hansen and Jaszkiewicz (1998) and *ONVG ratio* Van Veldhuizen (1999)), or to the size of the solution set itself (e.g., $C2$ Hansen and Jaszkiewicz (1998) and *error ratio* Van Veldhuizen (1999)). Beside, there also exist other indicators that simply count the number of solutions that are members of the optimal set Sierra and Coello (2005).

Since cardinality of a solution set usually has little information relevant to the representativeness of the Pareto front, it is often regarded less important than the other three quality aspects. However, evaluating the cardinality quality may become more reasonable if an optimiser can find a significant percentage of the problem's Pareto optimal solutions. This is particularly true for some combinatorial multi-objective optimisation problems where the total number of the Pareto optimal solutions is small. In such problems, counting the number of the obtained Pareto optimal solutions is a reliable indicator reflecting the solution set quality. In fact, this evaluation has been frequently used in some early studies on combinatorial problems, for example, see Ishibuchi and Murata (1998).

3.4.1 Error Ratio (ER)

The ER indicator Van Veldhuizen (1999) considers the proportion of nondominated solutions of the solution set that are the Pareto optimal solutions. Mathematically,

$$ER(\mathbf{A}) = \frac{\sum_{\mathbf{a}\in\mathbf{A}} e(\mathbf{a})}{N} \tag{5}$$

where N is the size of set $\mathbf{A}$, and

$$e(\mathbf{a}) = \begin{cases} 0, & \text{if } \mathbf{a} \in \text{PF} \\ 1, & \text{otherwise} \end{cases}$$

A smaller ER value is preferable. An ER value of zero indicates that every solution in the considered set is the Pareto optimal solution of the problem. However, since ER only counts in the Pareto optimal solutions, it may bring about some counter-intuitive cases. For example, adding more nondominated solutions to a set may lead to a worse ER score. As such, considering the nondominated solutions in the compared sets themselves (e.g., counting the unique nondominated solutions Berry and Vamplew (2005)) may probably be a better option, and also there is no need for the Pareto front.

3.5 QIs for Spread and Uniformity

As stated before, the quality aspects of spread and uniformity are closely related, and they need to be considered together to reflect the diversity of solution sets. This inspires QIs to cover both the spread and uniformity quality. Most QIs in this category can be put into two classes, distance-based indicators and region division-based indicators, despite the existence of alternatives, like the cluster-based indicator Li et al. (2005); Meunier et al. (2000) and the volume-based indicator Jiang et al. (2016); Tian et al. (2019). Table 2 lists 18 QIs for spread and uniformity in the literature (items 55–74).

3.5.1 Distance-based QIs

QIs in this class (items 59–62 in the table) typically consider distances between a solution and its neighbours and then sum up these distances, so as to estimate the coverage of the whole set. The first QI along this line is Δ Deb et al. (2002), followed by *sparsity index* Nicolini (2004), *extended spread* Zhou et al. (2006), and Δ_{Line} Ibrahim et al. (2017). However, such evaluation can only work in bi-objective problems as where nondominated solutions are located consecutively on either objective. Another issue of these QIs is that they require information of the Pareto front (e.g., the boundaries) as a reference, which is often unknown in practice.

3.5.2 Region Division-based QIs

The basic idea of this class (items 63–74 in the table) is to divide a particular space into many equal-size cells (with overlapping or not), and then to count the number of cells having solutions of the set. This is based on the fact that a set of more diversified solutions usually occupy more cells. The majority of QIs for spread and uniformity falls into this class, taking into account different shapes of the cells. Some of them consider niche-like cells which are centred by solutions themselves, such as *Chi-square-like deviation* Srinivas and Deb (1994), *U-measure* Leung and Wang (2003) and *sparsity* Deb et al. (2005). Some consider grid-like cells which divide the space into many hyperboxes, such as *cover rate* Hiroyasu et al. (2000), *number of distinct choices* Wu and Azarm (2001), *diversity metric* Deb and Jain (2002), *entropy* Farhang-Mehr and Azarm (2003a) and *diversity comparison indicator* Li et al. (2014a). The rest considers fan-shaped cells which divide the space by a group of uniformly-distributed rays (i.e., weight vectors), such as *Sigma diversity metric* Mostaghim and Teich (2005), *M-DI* Asafuddoula et al. (2015) and *DIR* Cai et al. (2018). In addition, dividing the space via considering the minimum energy points (s-energy) Hardin and Saff (2004) is also a potential way as they can well represent the space with various shapes.

3.5.3 Diversity Comparison Indicator (DCI)

Quality indicators in this category may not be very pragmatic, especially in high-dimensional scenarios. In distance-based QIs, the information of the problem's Pareto front is required, and in region division-based QIs, the number of the considered cells typically increase exponentially with the objective dimensionality. However, DCI Li et al. (2014a) seems an exception on this point. It does not need problem information, and the computational complexity is also quadratic.

DCI evaluates the relative quality amongst several solution sets. To do so, first all the sets under consideration are placed into a grid so there are some boxes having one or several solutions. DCI considers the boxes where nondominated solutions of the combined collection of all the sets are located. Then for each of these boxes, DCI calculates the contribution degree of every set to it. The contribution degree is a value $\in [0, 1]$, decreasing monotonously with the increase of the distance from the set to the box. A value of one means that there exist at least one solution in the box, and zero means that there is no solution in the box's neighbourhood, where the neighbourhood size increases consistently with the number of objectives. Finally, DCI averages the contribution degree of the set to all the considered boxes as its evaluation result. It assigns n scores (in the range of $[0, 1]$) to n solution sets. A set will have a high score if its solutions cover or are close to all the boxes, and if its solutions are far away from most of the boxes, a low score will be obtained. However, it is worth noting that as the considered boxes are usually not distributed uniformly, DCI may prefer the set which has a similar distribution with others.

3.6 QIs for All Quality Aspects

Quality indicators in this category are most commonly used in the literature, as they cover all the four aspects of solution sets' quality. Items 75–100 in Table 2 list these QIs. They, in general, can be divided into two classes: distance-based QIs (items 75–91) and volume-based QIs (items 92–100).

3.6.1 Distance-based QIs

The basis idea of distance-based QIs is to measure the distance of the Pareto front to the solution set under consideration. As such, a reference set that well represents the Pareto front is required. Only the solution set that is close to every member of the reference set can have a good evaluation value, thus a reflection of all the quality aspects convergence, spread, uniformity, and cardinality. This idea can be materialised by averaging (or summing up) the distances of the reference set's members to

their closest solution in the solution set, or finding the maximum value from these distances. With the former, *inverted generational distance* (IGD) Coello and Sierra (2004) is a representative example, which considers the average Euclidean distance. Other examples include Dist1 ($D1$) Czyzak and Jaszkiewicz (1998) and some IGD's variants Bosman and Thierens (2005); Eskandari et al. (2007); Ibrahim et al. (2017); Tian et al. (2016). They use difference distance metrics (e.g., Tchebycheff distance Czyzak and Jaszkiewicz (1998) and Hausdorff distance Schutze et al. (2012)) or introduce the dominance relation Dilettoso et al. (2017); Ishibuchi et al. (2015) or additional points Tian et al. (2016) in the evaluation.

Measuring the maximum difference (distance) of the Pareto front to the solution set can easily identify the gap between them, thus telling whether the solution set has a good coverage over the front Hadka and Reed (2012). Dist2 ($D2$) Czyzak and Jaszkiewicz (1998) and ϵ-indicator Zitzler et al. (2003) are such QIs. The Dist2 indicator considers Tchebycheff distance, while the ϵ-indicator considers the maximum difference on the objectives where the point of the reference set is superior to the solution of the considered set. Unlike averaging difference-based QIs, maximum difference-based QIs may have clear physical meaning; for example, the ϵ-indicator is to measure the minimum value added to any solution in the set to make it weakly dominated by at least one point in the reference set. However, their results typically only involve one particular solution on one particular objective, thus naturally lots of information loss.

Very recently, a quality indicator, called dominance move (DoM) Li et al. (2019), has been presented and can be seen as a combination of the above two types of QIs. DoM considers the minimum move of one set needed to weakly Pareto-dominate the other set. Specifically, given two solution sets $\mathbf{A}$ and $\mathbf{B}$, the DoM of $\mathbf{A}$ to $\mathbf{B}$ is the minimum total distance of moving some points in $\mathbf{A}$ such that any point in $\mathbf{B}$ is weakly Pareto-dominated by at least one point in $\mathbf{A}$. This intuitive indicator has many desirable properties, e.g., a natural extension of the comparison between two solutions, compliance with Pareto dominance, and no need of problem knowledge and parameters. However, its calculation is not trivial. While an efficient calculation method in the bi-objective case has been presented Li et al. (2019) , how to efficiently calculate it in the case with three or more objectives remains to be explored. It is worth mentioning that an earlier quality indicator Li et al. (2015a) could be seen kind of a simplified version of DoM. It divides the reference set (i.e., $\mathbf{A} \cup \mathbf{B}$) into many clusters and then sums up the maximum difference of $\mathbf{A}$ to each cluster. This renders the calculation efficient, but naturally loses its physical meaning.

3.6.2 Volume-based QIs

QIs in this class (items 92–100 in Table 2) measure the size of the volume determined by the considered solution set in conjunction with some specifications. For example, the widely used indicator *hypervolume* (HV) Zitzler and Thiele (1998) calculates the volume of the area enclosed by the set and a reference point specified by the user. Later, the HV indicator was modified/extended by incorporating normalisation before the calculation, e.g., *hyperarea ratio* Van Veldhuizen (1999) and *hyperarea difference* Wu and Azarm (2001), or by considering the difference of the areas that two solution sets dominate, e.g., *coverage difference* Zitzler (1999) and *volume measure* Fieldsend et al. (2003). In addition, the QI *completeness indicator* Lotov et al. (2013, 2002), which calculates the probability that a randomly generated solution from the decision space is weakly dominated by the considered solution set, is closely related to HV. The difference between them is that the completeness indicator needs sampling in the decision space but not the reference point in the objective space.

Other volume-based QIs include the R family (i.e., $R1$, $R2$ and $R3$) Hansen and Jaszkiewicz (1998) and *integrated preference functional* (IPF) Carlyle et al. (2003) and its variants Bozkurt et al. (2010); Fowler et al. (2005); Kim et al. (2006). Conceptually, the R family and IPF are similar to each other; both integrate over a set of utility functions (aka scalarising functions) which are assumed to be in accordance with the DM' preferences. Difference between them is that IPF considers a

parameterised set of utility functions with respect to the continuous weight space, while the R family directly considers the integration of the utility functions, despite that it is implemented by a discrete, finite set of weights.

Finally, it is worth noting that as seen in Table 2, there is no QI that is able to well reflect all the four quality aspects. This should not be surprising, since quality aspects can be conflicting to each other to some extent, such as convergence versus uniformity. For example, consider a set of uniformly distributed solutions. One can move one solution in the set a little to make it have better convergence. This will result in a new set having a better convergence but worse uniformity. Apparently, it is impossible for a single QI to catch both quality aspects in this example. On another note, it can be seen from the table that only HV (and its variants, items 93–96 and 100) in this category is able to well reflect the quality cardinality. This is because HV is the only known indicator fully consistent with the "better" relation (see Section 2.1), and thus it is sensitive to any change of nondominated solutions in a set.

3.6.3 Inverted Generational Distance (IGD)

IGD Coello and Sierra (2004) is amongst the most commonly used indicators, despite some similar ideas presented earlier Bosman and Thierens (2003); Czyzak and Jaszkiewicz (1998); Ishibuchi et al. (2003). As the name suggested, IGD is an inversion of the GD indicator, namely, to measure the distance from the Pareto front to the solution set.

Formally, given a solution set $\mathbf{A}$ and a reference set $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_M\}$,

$$IGD(\mathbf{A}, \mathbf{R}) = \frac{1}{M} \sum_{i=1}^{M} \min_{\mathbf{a} \in \mathbf{A}} d_2(\mathbf{r}_i, \mathbf{a}) \tag{6}$$

where $d_2(\mathbf{r}_i, \mathbf{a})$ denotes the Euclidean distance between $\mathbf{r}_i$ and $\mathbf{a}$. A low IGD value is preferable, and it should indicate that the set has good combined quality of convergence, spread, uniformity and cardinality.

However, the accuracy of the IGD evaluation largely depends on the approximation quality of the reference set to the Pareto front. Different reference sets can make the indicator prefer different solution sets. Usually, a large reference set with high resolution to the Pareto front is suggested. As shown in Ishibuchi et al. (2018b, 2015), insufficient points in the set can easily lead to counter-intuitive evaluations. In addition, the reference set consisting of all nondominated solutions generated by the optimisers may also cause misleading results, although this practice has been widely adopted in real-world problems. This issue will be discussed in detail in Section 4.4.2.

3.6.4 ϵ-indicator

Unlike IGD, the ϵ-indicator Zitzler et al. (2003) considers the maximum difference between solution sets. It is inspired by ϵ-approximation, a well-known measure for designing and comparing approximation algorithms in optimisation, operational research and theoretical computer science Helbig and Pateva (1994); Papadimitriou and Yannakakis (2000); Vaz et al. (2015).

Given two solution sets, the ϵ-indicator is the minimum factor by which one set has to be translated in the objective (in the way of addition or multiplication) so as to weakly dominate the other set. This thereby leads to two versions: the additive ϵ-indicator and the multiplicative ϵ-indicator. Mathematically, the additive ϵ-indicator of a set $\mathbf{A}$ to a set $\mathbf{B}$ is defined as

$$\epsilon_+(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{b} \in \mathbf{B}} \min_{\mathbf{a} \in \mathbf{A}} \max_{j \in \{1...m\}} \mathbf{a}_j - \mathbf{b}_j \tag{7}$$

where $\mathbf{a}_j$ denotes the objective of $\mathbf{a}$ in the jth objective and m is the number of objectives. The

multiplicative ϵ-indicator of $\mathbf{A}$ to $\mathbf{B}$ is defined as

$$\epsilon_{\times}(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{b} \in \mathbf{B}} \min_{\mathbf{a} \in \mathbf{A}} \max_{j \in \{1 \ldots m\}} \frac{\mathbf{a}_j}{\mathbf{b}_j} \tag{8}$$

Both indicators are to be minimised. A value of $\epsilon_{+}(\mathbf{A}, \mathbf{B}) \leq 0$ or $\epsilon_{\times}(\mathbf{A}, \mathbf{B}) \leq 1$ implies that $\mathbf{A}$ weakly dominates $\mathbf{B}$. When replacing $\mathbf{B}$ with a reference set $\mathbf{R}$ that represents the Pareto front, the ϵ-indicator can be used as a unary indicator. It measures the gap of the considered set to the Pareto front. However, as the returned value only involves one particular objective of one particular solution in either set (where the maximum difference is), the indicator may omit a significant amount of sets' difference. This can lead to differently-performed solution sets having the same/similar evaluation results, as observed in Liefooghe and Derbel (2016).

3.6.5 Hypervolume (HV)

HV Zitzler and Thiele (1998) was first presented as the *size of the space covered*, and later used as several terms *hyperarea metric* Van Veldhuizen (1999), *S metric* Zitzler (1999), and *Lebesgue measure* Fleischer (2003). The HV indicator is arguably the most commonly used QIs, due to its desirable practical usability and theoretical properties. Calculating HV does not need a reference set of representing the Pareto front, which makes it suitable for many real-world optimisation scenarios. The HV result is sensitive to any improvement to a set with respect to Pareto dominance. Whenever a set $\mathbf{A}$ is better than another set $\mathbf{B}$ (i.e., $\mathbf{A} \lhd \mathbf{B}$), then HV returns $\mathbf{A}$ a higher quality value than $\mathbf{B}$. Consequently, a set that achieves the maximum HV value for a given problem will contain all Pareto optimal solutions.

The HV indicator can be defined as follows. Given a solution set $\mathbf{A}$ and a reference point $\mathbf{r}$, HV can be calculated as

$$HV(\mathbf{A}) = \lambda\Big(\bigcup_{\mathbf{a} \in \mathbf{A}} \{\mathbf{x} | \mathbf{a} \prec \mathbf{x} \prec \mathbf{r}\}\Big) \tag{9}$$

where λ denotes the Lebesgue measure. Put it simply, the HV value of a set can be seen as the volume of the union of the hypercubes determined by each of its solutions and the reference point (as the left-bottom vertex and the right-top vertex, respectively).

A limitation of the HV indicator is its exponentially increasing runtime with regards to the number of objectives (unless $P = NP$) Bringmann and Friedrich (2010a). A review of the HV computational complexity will be given later (Section 4.2). Another issue of the HV indicator is the setting of its reference point. There is still no consensus on how to choose a proper reference point for a given problem, despite some common practices, e.g., the nadir point of the Pareto front or 1.1 times the nadir point of the compared solution sets' collection. Different reference points can lead to inconsistent HV evaluation results Knowles and Corne (2003). There is a lack of systematic studies/theoretical guidelines on the choice of the reference point in HV, except for a few in particular situations Auger et al. (2009b); Cao et al. (2015). Recently, Ishibuchi et al. (2017, 2018a) have demonstrated a clear difference of specifying the proper reference point for problems with a simplex-like Pareto front and an inverted simplex-like Pareto front. They have also shown experimentally that a slightly worse reference point than the nadir point is not always appropriate especially for the case of many-objective optimisation and/or a small population size. In addition, the HV indicator prefers the knee regions, and is biased to convex regions over concave regions Zitzler and Thiele (1998). As proven in Auger et al. (2009b), the distribution of a set of solutions that achieves the maximum HV value depends largely on the slope of the Pareto front. For example, HV may be in favour of very non-uniform solution sets on a highly non-linear Pareto front. This has been shown in Li et al. (2015a).

3.6.6 Integrated Preference Functional (IPF)

As a group of well-established QIs in operational research, IPF Carlyle et al. (2003) measures the volume of the polytopes determined by each of nondominated solutions in a set and a given utility function over the corresponding optimal weights. It can be perceived as representing the expected utility that a solution set carries for the DM Bozkurt et al. (2010). The IPF indicator is calculated by two steps: 1) to find the optimal weight interval for each nondominated solution and 2) to integrate the utility function over these optimal weight intervals.

Formally, let $\mathbf{A} \subset \mathbb{R}^m$ be a set of nondominated solutions, where m denotes the number of objectives. Consider a parameterised family of utility functions $u(\mathbf{a}, w)$ in which a given weight w produces a value function to be optimised, where $\mathbf{a} \in \mathbf{A}$ and $w \in W \subset \mathbb{R}^m$. For a given w, let $u^*(\mathbf{A}, w)$ be the best utility function value of the solutions in $\mathbf{A}$. Given a weight density function $h : W \longrightarrow \mathbb{R}^+$ that stands for the probability distribution of the (unknown) weight w and it holds that $\int_{w \in W} h(w)dw = 1$, then the IPF value of the set $\mathbf{A}$ is

$$IPF(\mathbf{A}) = \int_{w \in W} h(w)u^*(\mathbf{A}, w)dw \tag{10}$$

The utility functions can be represented as a convex combination of objectives Carlyle et al. (2003) (i.e., the weighted linear sum function) or the weighted Tchebycheff function Bozkurt et al. (2010). The former takes only the supported solutions into account, while the latter covers all nondominated solutions. The IPF indicator can be used in/without the presence of the DM's input. When the preferences of the DM can be expressed in accordance with some partial weight space, IPF measures how well the set represents the preferred potions of the Pareto front. When there is no preference information available, where all weights can be assumed to occur equally (i.e., $h(w) = 1, \forall w \in W$), IPF measures how well the set represents the whole Pareto front. A lower IPF value is preferable. However, a limitation of using the IPF indicator is that its computational complexity increases exponentially with the number of objectives Bozkurt et al. (2010); Kim et al. (2006), as it needs to integrate over the (continuous) weight space.

3.6.7 R Family

Similar to the IPF indicator, the R family Hansen and Jaszkiewicz (1998) also integrates the DM's preferences into the evaluation. However, different from IPF, the integration in the R quality indicator is based on utility functions (rather than on the weight space). Given two solution sets $\mathbf{A}$ and $\mathbf{B}$, a utility function space U and a utility density function $h(u)$, it can be defined as

$$R(\mathbf{A}, \mathbf{B}, U) = \int_{u \in U} h(u)x(\mathbf{A}, \mathbf{B}, u)du \tag{11}$$

Depending on the outcome function $x(\mathbf{A}, \mathbf{B}, u)$, the R family has three indicators. $R1$ considers the probability of one preferred by the DM over the other, $R2$ takes into account the expected values of the utility functions (which is like the IPF indicator), and $R3$ introduces the ratio based on $R2$. Among them, $R2$ is most frequently used, and can be expressed as

$$R2(\mathbf{A}, \mathbf{B}, U) = \int_{u \in U} h(u)u^*(\mathbf{A})du - \int_{u \in U} h(u)u^*(\mathbf{B})du \tag{12}$$

where $u^*(\mathbf{A})$ means the best value achieved by $\mathbf{A}$ on this specific utility function. As can be seen, the $R2$ value of two sets can be calculated separately. Like in the IPF indicator, $h(u)$ can be uniformly distributed over U, when the preference information unavailable. However, a discrete and finite set U is typically employed in the calculation, which is in contrast to a continuous set W considered in

IPF. This can make $R2$ computation friendly. In particular, if the set of utility functions u can be represented by a set of weights W and a parameterised utility function on these weights, then $R2$ can further be calculated as

$$R2(\mathbf{A}) = \frac{1}{|W|} \sum_{w \in W} u^*(\mathbf{A}, w) \tag{13}$$

Like in IPF, multiple choices exist in materialising $u(\mathbf{A}, w)$, such as the weighted linear sum function and the weighted Tchebycheff function, though the latter is widely used in practice Brockhoff et al. (2012, 2015); Zitzler et al. (2008).

Comparing Equation (13) with Equation (10) when $h(w)$ is set to 1, the $R2$ and IPF indicators appear quite similar. The IPF indicator, which considers a continuous weight space, requires exponentially growing computational time with respect to objective dimensionality, while the $R2$ indicator, which considers a discrete set of weights, is calculated quickly but naturally has a lower accuracy than IPF.

4 Important Issues in Quality Evaluation

In this section, we discuss several important issues about quality evaluation. It consists of attributes that QIs possess (Sections 4.1–4.2) and properties that QIs are desirable to have (Sections 4.3–4.6). Table 3 summarises these attributes and desirabilities with respect to the example QIs detailed in the last section.

4.1 Unary, Binary or M-nary indicators

The number of solution sets that QIs handle can be different. Although most QIs are unary indicators which independently evaluate a solution set by assigning it a real number, there exist some M-nary indicators ($M \geqslant 2$) which give relative quality of M solution sets by typically assigning them M real numbers. For example, the R family Hansen and Jaszkiewicz (1998) and the $\mathcal{C}$ indicator Zitzler and Thiele (1998) are amongst the earliest binary QIs, followed by the ϵ-indicator Zitzler et al. (2003), while G-metric Lizárraga-Lizárraga et al. (2008a), DCI Li et al. (2014a), and PCI Li et al. (2015a) were designed to compare any number of sets at one run. Compared with unary QIs, M-nary QIs have some strengths, such as less reference information required (as the considered sets can mutually be referred to each other), and more easily compliant with Pareto dominance (if introducing the comparison of the dominance relation amongst the sets in the evaluation).

However, as pointed out in Knowles and Corne (2002), M-nary indicators may induce cyclic relationship of the evaluation results. This contrasts with the fact that a QI is expected to have the transitivity property in the sense of providing a complete order of the solution sets being compared. For example, consider three bi-dimensional solution sets $\mathbf{A} = \{(2, 2), (0, 4)\}$, $\mathbf{B} = \{(3, 1.1), (0.8, 3.2)\}$, and $\mathbf{C} = \{(3.4, 0.8), (1, 3)\}$. The additive ϵ-indicator evaluates $\mathbf{A}$ better than $\mathbf{B}$ ($\epsilon_+(\mathbf{A}, \mathbf{B}) = 0.9 < \epsilon_+(\mathbf{B}, \mathbf{A}) = 1$), $\mathbf{B}$ better than $\mathbf{C}$ ($\epsilon_+(\mathbf{B}, \mathbf{C}) = 0.3 < \epsilon_+(\mathbf{C}, \mathbf{B}) = 0.4$), but $\mathbf{A}$ worse than $\mathbf{C}$ ($\epsilon_+(\mathbf{A}, \mathbf{C}) = 1.2 > \epsilon_+(\mathbf{C}, \mathbf{A}) = 1$). Other binary relations, e.g., symmetry, asymmetry and antisymmetry, are usually not satisfied in binary QIs as well. In addition, regarding the property of triangle inequality, it holds for some QIs, e.g., the ϵ-indicator, but not for some others, e.g., the $\mathcal{C}$ indicator. For the latter, taking three bi-dimensional sets $\mathbf{A} = \{(1, 1)\}$, $\mathbf{B} = \{(0, 3)\}$, and $\mathbf{C} = \{(2, 2)\}$ as an example, we have that $\mathcal{C}(\mathbf{A}, \mathbf{B}) + \mathcal{C}(\mathbf{B}, \mathbf{C}) = 0 + 0 = 0 < \mathcal{C}(\mathbf{A}, \mathbf{C}) = 1$.

It is worth mentioning that some M-nary QIs can be converted into unary ones. When evaluating solution sets, if an M-nary QI is modified not to compare them with each other, but to compare them with the Pareto front of the problem or the collection of all the sets under consideration. This leads to a unary indicator, despite that this conversion does not work for all M-nary QIs, e.g., the $\mathcal{C}$ indicator. The resulting unary QI now has the transitivity property and induces a complete order of

Table 3: Attributes and desirabilities of the example quality indicators in Section 3. The symbol "√" in the last four columns means the QI has the specified desirability.

Indicator	Number of sets	Computational effort	Pareto compliant	Additional problem knowledge	No need of scaling before calculation	Effect of dominated or duplicate solutions
$\mathcal{C}$ indicator	binary	quadratic	√	√	√	both
GD	unary	quadratic		Pareto front		both
MS	unary	linear		√		dominated
SP	unary	quadratic		√		both
ER	unary	quadratic		Pareto front	√	both
DCI	arbitrary	quadratic	√[a]	grid division	√	dominated
IGD	unary	quadratic		reference solution set		dominated
ϵ-indicator	unary/binary	quadratic	√	√ for binary ϵ		√
HV	unary	exponential in m	strictly √	reference point	√	√
IPF	unary	exponential in m	√	reference weight set, (possibly) reference point		√
$R2$	unary/binary	quadratic	√	reference weight set, reference point		√

[a]DCI is Pareto compliant when comparing two solution sets, but not if more sets are involved.

solution sets. On the other hand, some unary QIs can be converted into binary QIs. For example, HV Zitzler and Thiele (1998) can be easily transformed into a binary indicator as done in Zitzler and Thiele (1999) (as the coverage difference indicator), where the areas that one set dominates but the other not are considered.

4.2 Computational Effort

Most QIs are cheap to compute. The time complexity of cardinality QIs and diversity QIs (i.e., evaluating spread and/or uniformity) is typically linear and quadratic, respectively, in the size of the solution set. QIs involving a reference solution/weight set often require $O(mNR)$ comparisons, e.g., the GD, IGD and $R2$ indicators, where m denotes the number of objectives, N the size of the solution set, and R the size of the reference set; thus, no need to worry about the running time of such QIs under the reasonable size of the reference set. For M-nary QIs, $O(mN^2M)$ comparisons are usually required as they need to compare each solution in a set to all solutions of other sets, such as the $\mathcal{C}$, DCI and ϵ indicators. HV, IPF and their variants are amongst expensive QIs, with their computational time increasing exponentially with the number of objectives. This may limit their application in the high-dimensional space, especially as an indicator integrated into the search process of an optimiser. The most costly QI is DoM Li et al. (2019) whose computational time known increases exponentially with the size of the solution set.

Computing the HV indicator has received intensive attention in the last 15 years. This problem was shown to be a special case of Klee's measure problem Beume (2009). For the 2D and 3D case, the optimal computational time is $O(N \log N)$ Beume et al. (2009). To reduce the theoretical time complexity in the general case, lots of attempts have been made, see for example Beume et al. (2009); Bringmann (2013); Bringmann and Friedrich (2010b); Fonseca et al. (2006); While et al. (2006); Yildiz and Suri (2012). Among them, the best known algorithm achieves a $O(N^{m/3}\text{polylog}N)$ upper bound Chan (2013), where N denotes the size of the set and m denotes the number of objectives. However, there seems no available implementation of this method and no evidence of its practical efficiency. From a practical viewpoint, methods in Guerreiro et al. (2012); Jaszkiewicz (2018); Lacour et al. (2017); Russo and Francisco (2014); While et al. (2012) are amongst the most efficient ones. These efforts make the HV indicator workable in evaluating solution sets (with a reasonable size) having more than 10 objectives, but may still struggle when used as an indicator to guide the search

in an optimiser. Although approximating HV by the Monte Carlo sampling can significantly reduce its calculation time Bader and Zitzler (2011); Bringmann et al. (2013); Ishibuchi et al. (2010), it is still faced the issue of "curse of dimensionality" — it may fail to distinguish between different solution sets when the number of objectives reaches up to 20.

The space requirement of QIs is usually minor, with the exception of those using the grid in the evaluations, such as DM Deb and Jain (2002) and the entropy indicator Farhang-Mehr and Azarm (2003a). In these QIs, one needs to access each box in the grid to record its information; for a solution set with m dimensions, d^m boxes need to be considered, where d is the number of divisions in each dimension. This problem, though, seems not intractable if only considering the non-empty boxes instead (i.e., the boxes have at least one solution of the set) Yang et al. (2013). In this case, there are at most N boxes whose information needs to be stored.

4.3 Pareto Compliance

A quality indicator is said to be *Pareto compliant* Knowles et al. (2006); Zitzler et al. (2003) (or *monotonic* Zitzler et al. (2008)) if and only if, when comparing one solution set with another, "at least as good" in terms of the dominance relation implies "at least as good" in terms of the QI values Zitzler et al. (2008). Formally, it can be expressed as follows (assuming that a smaller evaluation result is preferable):

$$\forall \mathbf{A}, \mathbf{B} : \mathbf{A} \preceq \mathbf{B} \Rightarrow I(\mathbf{A}) \leq I(\mathbf{B}) \tag{14}$$

For M-nary QIs, $I(\mathbf{A})$ can be seen as the relative quality of the set $\mathbf{A}$ in comparison with the other set(s). Pareto compliance guarantees that a QI does not contradict the order of solution sets induced by the weak Pareto dominance relation (also by the other comparison relations introduced in Section 2.1 as the weak Pareto dominance relation is the weakest out of them), i.e., if set $\mathbf{A}$ is evaluated better than set $\mathbf{B}$, it would never happen that $\mathbf{B}$ weakly Pareto dominates $\mathbf{A}$.

Many widely-used QIs are Pareto non-compliant, such as GD, SP, MS, and IGD. It has been frequently shown that they may violate the partial order of Pareto dominance amongst solution sets Ishibuchi et al. (2015); Knowles and Corne (2002); Knowles et al. (2006); Schutze et al. (2012); Zitzler et al. (2003). This is particularly true for QIs that measure the diversity (i.e., spread and/or uniformity) of solution sets, since they rarely take into account the closeness of solution sets. DCI is the only known diversity QI that is compliant with Pareto dominance when comparing two sets. But if more sets are involved, DCI does not guarantee the Pareto compliance property. Some Pareto non-compliant QIs involving convergence evaluation can be compliant under some specific situations, such as GD and IGD when used for a bi-objective problem with a continuous Pareto front Schutze et al. (2012). On the other hand, some non-compliant QIs, after modifications, can become Pareto compliant. The studies in Ishibuchi et al. (2015) are an example along these lines, where GD and IGD are transformed into two Pareto compliant indicators GD^+ and IGD^+, respectively. It is worth mentioning that M-nary indicators seem more easy to be transformed, as they can introduce the dominance relation of the compared solution sets first in the evaluation.

Note that Pareto compliant QIs may not fully distinguish between solution sets subject to the "better" relation, namely, the most general form of superiority of solution sets. That is, when $\mathbf{A} \lhd \mathbf{B}$, a Pareto compliant indicator I may return $I(\mathbf{A}) = I(\mathbf{B})$. To deal with this, a stronger condition is introduced in Zitzler et al. (2007), called strict Pareto compliance:

$$\forall \mathbf{A}, \mathbf{B} : \mathbf{A} \lhd \mathbf{B} \Rightarrow I(\mathbf{A}) < I(\mathbf{B}) \tag{15}$$

The strict Pareto compliance implies that only the Pareto front achieves a unique optimal value for a problem. So far, the HV indicator (and its variants) is the only popular unary QI having this

property.[3]

4.4 Additional Problem Knowledge

Most QIs require additional problem information, especially for convergence-related QIs which typically need some sort of reference of the problem's Pareto front for a comparison, e.g., the ideal point, the nadir point, or a reference set that represents the front. As the accuracy of QIs can be largely dependent on such references, it is desirable for QIs to have as little reference information as possible. When reporting the results of a study, it is strongly advised to accompany the evaluation values with the reference information used, for example, by explicitly noting the reference point used or making the reference set used public, so that others can stand on the same page to compare their results.

4.4.1 Reference Point

The frequently-used reference points in QIs are the ideal point and the nadir point (or the one derived from them). For example, Tchebycheff utility function-based IPF and $R2$ need the ideal point of the problem and the HV indicator needs a reference point that is typically derived from the nadir point. Although the true ideal point of the problem is usually unavailable, we may use an estimated point obtained by the solution sets under consideration, or by separately optimising each of the objectives of the problem. Note that it is crucial to not underestimate the ideal point as this may lead to solutions that are non-dominated with the estimated ideal point not to contribute fully to the evaluation results. Furthermore, even if an accurate ideal point is used, boundary solutions may still contribute less than the inner ones, such as in IPF and $R2$. As such, it is advised to use a reference point slightly better than the ideal point, say $ref_j = ideal_j - l_j/(2N-2)$ in a bi-objective scenario, where l_j is the range of the objective j and N is the size of the solution set. This setting is able to make the boundary solutions contribute (approximately) equally as the other internal solutions in the calculation of the concerned indicator (when considering a uniformly distributed solution set which spreads over a linear Pareto front).

Compared to the ideal point, determining the nadir point can be very difficult, even for simple problems Hansen and Jaszkiewicz (1998). One important reason for this is the presence of the dominance resistant solutions (DRSs) in a solution set Ikeda et al. (2001), i.e., the nondominated solutions with an extremely poor value in one objective but with (near) optimal values in some other objectives. If one estimates the nadir point by the collection of the solution sets under consideration in which there exist DRSs, the evaluation results could be largely affected. For example, in HV, setting the reference point equal to (or worse than) such DRSs on the corresponding objective may make some outer solutions contribute significantly more than the inner ones.

4.4.2 Reference Set

A reference set representing the Pareto front is needed in many QIs, such as IGD, the unary ϵ-indicator, and the relative HV (i.e., *hyperarea ratio* Van Veldhuizen (1999)). Ideally, such a reference set is expected to consist of sufficient points that are distributed uniformly and densely on the Pareto front.[4] This, though, is not feasible in most cases. A practical alternative is to use the collection of the solution sets under consideration as the reference set. However, this common practice may come with two issues. First, whenever new solution sets are included in the comparison, the reference set needs to be reconstructed. And more importantly, such a collection of the sets may not well represent the Pareto front, and consequently the QI could return inaccurate results.

[3] The *potential function* in Laumanns and Zenklusen (2011) also possesses this property, but its calculation involves the whole search space.

[4] A recent paper has shown that a well-distributed reference set could lead to IGD to prefer the internal points, and suggested an extension of the Pareto front as the reference set used in the evaluation Ishibuchi et al. (2018b).

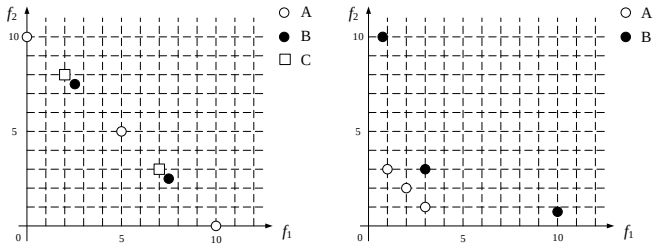

Figure 3: Illustrations that the collection of solution sets as the reference set may lead to misleading IGD results. (a) For three bi-dimensional sets $\mathbf{A} = \{(0, 10), (5, 5), (0, 10)\}$, $\mathbf{B} = \{(2.5, 7.5), (7.5, 2.5)\}$ and $\mathbf{C} = \{(2, 8), (7, 3)\}$, in general $\mathbf{A}$ can be seen to outperform $\mathbf{B}$ and $\mathbf{C}$ as it provides better spread and cardinality, but IGD gives opposite results: $\text{IGD}(\mathbf{A}) \approx 1.82 > \text{IGD}(\mathbf{B}) \approx 1.72 > \text{IGD}(\mathbf{C}) \approx 1.61$. (b) For two bi-dimensional sets $\mathbf{A} = \{(1, 3), (2, 2), (3, 1)\}$ and $\mathbf{B} = \{(0.75, 10), (3, 3), (10, 0.75)\}$, the two boundary points of $\mathbf{B}$ can be seen as the dominance resistant solutions Ikeda et al. (2001) (i.e, slightly better on one objective but significantly worse on the other objective), thus unlikely to be preferred by the DM, but IGD gives opposite results: $\text{IGD}(\mathbf{A}) \approx 2.80 > \text{IGD}(\mathbf{B}) \approx 1.08$.

On one hand, the reference set may make QIs prefer a specific distribution pattern consistent with the majority of the considered solution sets. This means that if one solution set is distributed very differently from others, then this set is likely to assign a poor evaluation value whatever its actual distribution is. This particularly applies to a QI that is sensitive to the reference set's distribution, such as IGD and IGD^+. Figure 3(a) gives such an example with respect to IGD. When comparing $\mathbf{A}$ with $\mathbf{B}$, if the reference set is composed of these two sets, we will have $\mathbf{A}$ evaluated better than $\mathbf{B}$ ($\text{IGD}(\mathbf{A}) \approx 1.41 < \text{IGD}(\mathbf{B}) \approx 2.12$). But if adding another set $\mathbf{C}$ which has the similar distribution pattern to $\mathbf{B}$ into the evaluation, and now the reference set is composed of the three sets, we will have $\mathbf{A}$ worse than $\mathbf{B}$ ($\text{IGD}(\mathbf{A}) \approx 1.82 > \text{IGD}(\mathbf{B}) \approx 1.72$). A potential method to deal with this issue is to cluster crowded points in the reference set first and then to consider these well-distributed clusters instead of arbitrarily-distributed points, as done in the PCI indicator Li et al. (2015a). But this induces another issue: how to properly cluster the points in the reference set with potentially highly irregular distribution.

On the other hand, as discussed previously, the dominance resistant solutions (DRSs) can affect the evaluation results if they are contained in the reference set. Take Figure 3(b) as an example, where the two boundary solutions of the set $\mathbf{B}$ can be seen as the DRSs. The IGD indicator evaluates $\mathbf{B}$ better than $\mathbf{A}$, despite the fact that $\mathbf{A}$ converges significantly better than $\mathbf{B}$ and is likely to be preferred by the DM. This issue can be well overcome if only taking into account the superiority between two points when measuring their distance Li et al. (2014b), which was done in IGD^+ and PCI.

Overall, the reference set consisting of points with an arbitrary distribution and location may bring inconsistent evaluation results. But, the extent of this effect could be different to QIs. A QI, which is little sensitive to the distribution and DRSs of the reference set, can always accurately reflect the quality of the solution set, such as the ϵ-indicator and PCI which are able to work with the reference set having any distribution and location. The relative HV has no requirement of the reference set having uniformly distributed points, while IGD^+ is insensitive to the DRSs.

In addition, it is worth noting that other kind of reference sets may need in QIs, such as a set of reference weights used in the IPF and $R2$ indicators. Similar to the situation of the reference point set, the evaluation results are affected by the distribution of the reference weight set. In the implementation, typically a set of uniformly distributed weights in a unit simplex are considered. This implies a risk to perform the evaluation for problems without a simplex-like Pareto front. In which case, the QIs may prefer a set having specific distribution of solutions (e.g., more solutions concentrating around the boundaries of the set for problems with an inverted simplex Pareto front) to one having uniformly distributed solutions.

4.4.3 Parameters

Some QIs require setting parameters in the evaluation, e.g., the size of the niche in niche-related indicators and the number of divisions in grid-based ones. These parameters can be affected by the size of the considered solution set, the number of objectives, and also the actual dimensionality of the Pareto front. For the last one, it is hard to estimate as the manifold of different solution sets can be diverse. In addition, the "sweet-spot" of indicator parameter settings that produces reliable results could shrink significantly as the number of objectives increases. Two high-dimensional solution sets easily return opposite results when evaluated by a QI with slightly different settings of its parameter.

4.5 Scaling and Normalisation

Scaling (or normalisation) may need for QIs whose calculation involves objective blending. When the objectives are incommensurable, usually the bigger the range of the objective, the bigger the effect of that objective to the indicator values. This is opposite to the fact that typically we hope that different objectives contribute equally to the indicator values. So, a proper transformation of the objectives is unavoidable, in order to enable their values to lie in the (approximately) same range, e.g., $[0, 1]$.

A standard objective transformation form is $\mathbf{a'}_j = (\mathbf{a}_j - min_j)/(max_j - min_j)$, where min_j and max_j are the minimum and maximum respectively on the objective j with respect to the problem's Pareto front ideally, or with respect to the nondominated set of the collection of all the considered solution sets practically. However, this can also be affected by the DRSs. In this case, most solutions could be transformed into a tiny portion of the normalised range (e.g., $[0, \delta]$, where δ is a very small decimal) on some of the objectives, and consequently these objectives contribute less to the indicator values than the other objectives. A potential method to tackle this is to use nonlinear transformations by, for example, taking into account the dispersion of the solutions on the objectives. In addition, other ways to determine the range of the scaling include using the anti-ideal point, the probabilistic approach and the pay-off table, and adding "common sense" to these methods, which fit in particular situations Hansen and Jaszkiewicz (1998).

Note that one may not need to perform the normalisation if the concerned QI is *scaling independent* Zitzler et al. (2008). An indicator is called scaling independent if the order of solution sets induced by the indicator always remains the same after applying any monotonic transformation to the objective values Zitzler et al. (2008). But this desirable property only holds for QIs having no "blending" of the objectives, e.g., those dominance-based QIs and cardinality-based QIs.

It is worth pointing out that blending objectives in quality evaluation does not certainly need the normalisation, since different objectives may still contribute equally to the indicator values after blending. For example, in the HV calculation objectives need a "multiplication". Objectives with different ranges contribute proportionally the same amount to the volume. As a result, the ratio of HV before and after the normalisation remains unchanged for different solution sets (provided that the choice of the reference point takes the objective ranges into consideration). This observation is actually not in line with the common practice in the area, which normalises the solution set before calculating its HV value.

4.6 Effect of Adding Dominated or Duplicate Solutions

In general, it is desirable for a QI that adding dominated or duplicate solutions into a solution set does not affect the indicator results as these solutions bring no more useful information to the DM in the context of Pareto optimality. However, this does not hold for many QIs, particularly for those taking into account the contribution of each solution in the set to the indicator values, e.g., many convergence QIs and most diversity QIs. For cardinality QIs, adding duplicate solutions often has a problem as duplicate solutions are conceptually nondominated with the existing ones. For QIs

covering all the aspects of set quality, adding dominated solutions may affect the indicator values, such as IGD, but many Pareto compliant QIs possess this desirability in which dominated solutions contribute nothing to the indicator values, such as HV, IGD^+, IPF, PCI, $R2$, and the ϵ-indicator.

5 Future Research Directions

After providing a general review of the research currently done and important issues in quality evaluation, this section is devoted to indicating what are some of the promising research directions for the next few years, and also to giving some guidelines on these directions. This section focuses on major topics and some other interesting open problems can also be found in `http://simco.gforge.inria.fr/doku.php?id=openproblems`, such as parameter sensitivity, k-replacement landscape and set-gradient.

5.1 QI Design

An ideal quality indicator is cheap to compute, Pareto compliant, immune to dominated/duplicate solutions, does not need a normalisation and not need any additional problem knowledge. However, as can be seen in the last two sections, such an indicator does not exist currently. Even only for the two desirabilities Pareto compliance and additional problem knowledge, there is no unary QI both possessing them. All the unary Pareto compliant QIs (e.g., HV and the ϵ-indicator) require a reference point/set, while all the unary diversity QIs which typically need less parameter information (e.g., MS and SP) are not compliant with Pareto dominance. In this regard, M-nary QIs seem to be a better option, since it is easier for them to be Pareto compliant and the sets under consideration can be mutually used as the reference to each other. In addition, we would like to note that if it is unavoidable to introduce reference information in designing QIs, it may be better for them to relate to the ideal point of the problem, than on the nadir point, than on a reference set, as they are of different achievable levels in practice.

It is worth noting that a caveat needs to keep in mind when designing uniformity QIs. In a high-dimensional space, the ratio between the distances of a solution to its nearest and farthest solutions approaches 1, i.e., the solutions essentially become uniformly distant from each other. This phenomenon can be observed for a variety of distance metrics, but it is more pronounced for the Euclidean distance than, say, the Manhattan distance. So if a QI considers the Euclidean distance in its evaluation, then the evaluation results of different solution sets may tend to be similar in a high-dimensional space. To alleviate this issue, the L^p norm metric with $p < 2$ can be introduced. In this regard, the Manhattan distance (i.e., $p = 1$) is a good option as it does not violate the triangle inequality.

5.2 QI Selection

An interesting phenomenon in QI selection is that in spite of the considerably large number of QIs currently available (see Table 2), many researchers adopt just a handful of them. Some QIs (e.g. HV) have a clear preference over others (e.g., the ϵ-indicator). We think that this may be attributed to several reasons. First, researchers in the field tend to work by analogy. An example in this regard is that the indicator GD, which actually has an obvious drawback (i.e., sensitive to the size of a solution set because of the "quadratic mean" considered, cf. Section 3.1.4), is, however, used widely. In contrast, the indicator $\mathcal{M}_1^*$ Zitzler et al. (2000) which is similar to GD but overcomes its drawback, is, however, used rarely. This interesting phenomenon results from "inertia"; people tend to use QIs which was (widely) used before, and may not be fully aware of their drawbacks. Second, some QIs are more practical/user-friendly. For example, the HV indicator does not need a reference set that estimates the Pareto front, while the ϵ-indicator needs one (when used as a unary indicator), despite

both being Pareto compliant. The binary ϵ-indicator, like any binary QI, does pairwise comparisons of two solution sets. For n solution sets, they need $\binom{n}{2}$ evaluations, more than unary QIs requiring n evaluations. Also, to compare results obtained by binary QIs, more complicate statistical testing is needed. The last reason is that researchers in different fields may prefer different QIs. For example, people in the operational research field tend to consider the indicator IPF and its variants Bozkurt et al. (2010); Carlyle et al. (2003), while people in the search-based software engineering field like to choose the contribution indicator Meunier et al. (2000) to compare sets (through the dominance relation of their solutions) obtained by different algorithms. Overall, these reasons altogether suggest the importance of understanding the properties and advantages/drawbacks of various QIs, from which the user could find her/his suited one(s).

In general, it is expected to select QIs which together are able to cover all the four quality aspects. Many papers, especially those appeared one decade ago, are used to consider multiple QIs, each of which is responsible for one specific aspect. For example, a common practice is using GD for convergence, MS and/or SP for diversity, and ER for cardinality Adra and Fleming (2011); Coello et al. (2004). In contrast, now there is a trend of using a comprehensive QI to evaluate all the quality aspects, such as HV and IGD. However, there is no clear evidence indicating this practice better than a combination of multiple isolated QIs, which suggests further studies needed. But anyway, no matter what kind of combination (or comprehensive indicator), one needs to ensure it sufficient to properly cover all the four aspects, in which case, MS or SP (or their combination) may not suffice for diversity, as they omit how the solution set spreads over the space.

Recently, some studies have paid attention to the ensembles of several QIs Yen and He (2014) and the hierarchical evaluation through multiple QIs Lizárraga-Lizárraga et al. (2008a); Zitzler et al. (2008), but this relies heavily on the diversity amongst the selected QIs. When all selected QIs have similar behaviours, e.g., all preferring boundaries of the Pareto front, being better with respect to one QI usually implies being better with respect to the others, thereby failing to make a reliable conclusion of solution sets' comprehensive quality. In addition, as the importance of different quality aspects may be different (generally in the order of convergence, spread, uniformity, and cardinality), when designing a QI ensemble strategy one needs to take this order into consideration.

The above discussions are based on generic multiobjective problems. For a particular class of problems, there could be preferred QIs. For example, many problems in data mining, such as the feature selection problem, could be regarded as combinatorial problems Mukhopadhyay et al. (2014a,b). As combinatorial problems often have a relatively small Pareto front, evaluating the cardinality quality of solution sets (against all the known Pareto optimal solutions) may be preferred. Besides, for some problems we know the range of the Pareto front, such as the optimal product selection problem in software product line Hierons et al. (2016); Xiang et al. (2018). In this case, the indicator HV is a good option to evaluate solution sets as the reference point is able to properly set.

5.3 Connection between QIs

As diverse QIs are desirable to together evaluate solution sets, it is of high importance to understand the connection between existing QIs. Such study should be stressed on QIs responsible for same quality aspect(s) (e.g., those responsible for all the four aspects), since QIs for different aspects are typically independent of each other. Recently, there have been a few attempts along these lines. Wessing and Naujoks (2010) analysed the correlation structure between HV, $R2$ and the ϵ-indicator, and have reported that HV and $R2$ are highly correlated. Similar observations have been found in Liefooghe and Derbel (2016), where the authors conducted an extensive empirical study of the correlation between four QIs. They have also found that IGD appears to behave more differently from the other Pareto compliant QIs. But we may not know whether this claim results from the Pareto non-compliance property of IGD or not. In addition, Jiang et al. (2014) investigated the difference in indicator values and have shown the inconsistent results obtained by HV and IGD. More recently,

Table 4: QIs used in indicator-based search methods.

Indicator(s)	indicator-based search methods
HV	IBEA Zitzler and Künzli (2004), SMS-EMOA Beume et al. (2007); Emmerich et al. (2005), MO-CMA-ES Igel et al. (2007), HypE Bader and Zitzler (2011), POSEA Yevseyeva et al. (2014)
ϵ-indicator	IBEA Zitzler and Künzli (2004), PBEA Thiele et al. (2009)
$R2$	R2-EMOA Brockhoff et al. (2015); Trautmann et al. (2013), R2-IBEA Phan and Suzuki (2013), R2-MOGA Diaz-Manriquez et al. (2013), MOMBI Hernández and Coello (2013), MOMBI-II Hernández and Coello (2015)
Δ_p indicator	DDE Villalobos and Coello (2012), AS-EMOA Rudolph et al. (2014)
IGD or its variants	IGD+-MOEA Lopez and Coello (2016), AR-MOEA Tian et al. (2017), MaOEA/IGD Sun et al. (2018)
SDE	I_{SDE^+} Pamulapati et al. (2018)
ϵ-indicator + SDE	SRA Li et al. (2016)
ϵ-indicator + SDE + PBI	K-MBFA Wang et al. (2018)
Indicator ensemble	BIBEA-P Phan et al. (2012)

Ravber et al. (2017) studied the difference among QIs by ranking several well-established optimisers and have reported that HV, IGD^+ and the ϵ-indicator do not perform significantly differently in the optimiser ranking.

The above studies, however, have not provided deep insight of how to select QIs properly. They did not touch on the issues like if it is possible for several specific QIs to (approximately) cover another QI, or suggesting a proper (minimal) set of QIs which have a good coverage of solution set quality. In addition, studies on theoretical connection between QIs are also highly desirable. An important step in this respect has been made in Bringmann and Friedrich (2010c), showing that HV approximates the (additive) ϵ-indicator with asymptotically optimal convergence speed, and a set that maximises the HV value is near optimal on the ϵ-indicator as well, with the "gap" diminishing as quickly as $O(1/N)$, where N denotes the set size.

5.4 Optimal Distribution

Given a quality indicator and a Pareto front, optimal μ-distribution refers to the geometrical properties of solution sets (with the size μ) that maximise the indicator. Determining the optimal distribution is an important research topic as it is of relevance to understanding the bias of QIs and also to knowing the limiting distribution achieved by multiobjective optimisers with a bounded size of populations.

Optimal distribution is typically considered in comprehensive QIs. It was first described in Auger et al. (2009b) for the HV indicator, where it has shown that in the 2D case, the optimal distribution of solutions is most dense as parts of the Pareto front where the angle of the tangent with the x-axis is $-45°$, and it decreases to 0 for angles close to $0°$ and $90°$. Later on, they have shown some preliminary results in the 3D case Auger et al. (2010). Other results include determining the optimal distributions for HV with α-approximation Friedrich et al. (2009) and for the $R2$ indicator Brockhoff et al. (2012) in certain condition. However, it remains widely unclear the dependency of the density on the curvature of the Pareto front when three or more dimensions are involved.

5.5 QI-based Search

A trend in multiobjective optimisation is to integrate QIs into optimisers themselves. As the outcome of optimisers is eventually evaluated by QIs, it is natural and logical to directly optimise the QI value during the search process, thus resulting in the search in the space of sets rather than in the space of solutions. Such indicator-based search, along with Pareto-based search and decomposition-based search, have become three mainstream techniques in evolutionary multiobjective optimisation. Table 4 summarises indicator-based search methods and their corresponding QIs.

Early indicator-based optimisers typically integrate the HV or ϵ-indicator into their search, such as IBEA Zitzler and Künzli (2004), SMS-EMOA Beume et al. (2007), MO-CMA-ES Igel et al. (2007), PBEA Thiele et al. (2009), and HypE Bader and Zitzler (2011), Recent indicator-based optimisers introduce other quality indicators, including the $R2$ indicator Brockhoff et al. (2015); Diaz-Manriquez et al. (2013); Hernández and Coello (2015); Phan and Suzuki (2013), averaged Hausdorff distance Δ_p Rudolph et al. (2016, 2014); Villalobos and Coello (2012), IGD (or its variants) Lopez and Coello (2016); Sun et al. (2018); Tian et al. (2017), among others Basseur and Burke (2007); Thiele (2015). These methods have been found to be particularly promising in problems with many objectives as they can provide sufficient selection pressure (against Pareto dominance), returning a scalar value for a solution set with any dimension.

However, such set-based selection, compared with solution-based selection, typically accompanies with high computational cost. Every time we remove one solution in the set, it potentially needs N evaluations to determine which one having the least contribution to the indicator. Another issue of indicator-based optimisers is that they of course rely on the accuracy and the behaviours of the considered QI. Since each indicator represents one particular preference structure, it seems promising to combine an indicator with the general Pareto dominance relation or to consider multiple indicators with different preference structures. The former has been demonstrated in Li et al. (2018b), and the latter has been practised in several recent studies Li et al. (2016); Phan et al. (2012); Wang et al. (2018).

5.6 Hypervolume-Related Issues

The HV indicator has long been one of the central topics in the evolutionary multiobjective optimisation area. There exist several open problems that constantly attract the attention of researchers, such as the computational complexity, incremental update in HV-based search and subset selection. Readers of interest could refer to the following two webpages in detail, `http://simco.gforge.inria.fr/doku.php?id=openproblems` and `http://www.hypervolume.org`.

5.6.1 Computational Complexity

As stated previously, the known upper bound runtime is $O(N^{m/3}\text{polylog}N)$, where N denotes the set size and m denotes the number of objectives Chan (2013). A lower bound of $\Omega(N \log N)$ was given in Beume et al. (2009). Now it still remains unknown to find better lower bound and upper bound when $m > 3$. Also algorithms with a small constant (hidden in the $O()$ notation) would be desirable, even in small dimensions. In addition, computing sizes of the facets of the dominated HV polytope is of interest as it is closely related to the computation of the HV contribution Emmerich and Deutz (2014).

5.6.2 Incremental Update

Incremental update of HV (also called computing HV contributions) is to measure how much the HV value of a set changes, when one solution is added into or removed from the set. It plays an important role in HV-based archive maintenance Knowles et al. (2003) and HV-based evolutionary search Emmerich et al. (2005). Computing the HV contributions exactly is #P-hard and approximating them is NP-hard Bringmann and Friedrich (2012). In the 2D and 3D cases, computing all HV contributions requires $O(N \log N)$ comparisons Emmerich and Fonseca (2011), and in the 4D case, it can be obtained in $O(N^2)$ comparisons Guerreiro and Fonseca (2018). The asymptotically fastest known algorithms in more dimensions were presented in Bringmann and Friedrich (2012); Igel et al. (2007). Also it has been pointed out that the computation of a single HV contribution in dimension d has at least the complexity of the computation of the HV indicator in $d-1$ Emmerich and Fonseca (2011). In addition, as shown in Hupkens and Emmerich (2013), updating all HV contributions in an

archive of N nondominated solutions can be achieved with an amortised time complexity of $\Theta \log N$ after adding or removing a single solution in the 2D case. For the 3D and 4D cases, algorithms are known with (amortised) runtime linear and quadratic in N, respectively Guerreiro and Fonseca (2018). However, similar to the computation of the HV indicator, the time complexity of computing HV contributions in more than 3 dimensions remains unknown.

5.6.3 Subset Selection

Finding the optimal K-size subset with respect to HV from an N-size solution set is of interest in multiobjective optimisation. It maximises the probability that the DM finds at least one solution in the subset acceptable Emmerich et al. (2015), and is also relevant for maintaining a high quality population/archive in the search. This problem is in general NP hard for 3 and more dimensions Rote et al. (2016). In the 2D case, algorithms of time complexity $O(N(K + \log N))$ were presented Bringmann et al. (2014); Kuhn et al. (2016), faster than a dynamic programming-based algorithm of $O(KN^2)$ Auger et al. (2009a). Recently, a fixed parameter approximation algorithm for the general case was given in Bringmann et al. (2017). However, it is still unknown whether there exist more efficient algorithms for 2 dimensions and whether the problem is $W[1]$ complete for more than 2 dimensions. In addition, the subset selection problem also applies to other quality indicators, see for example Vaz et al. (2013) where finding a subset with respect to the best ϵ value was considered.

It is worth mentioning that recently some work Qian et al. (2018, 2016, 2015) converts a general subset selection problem into a bi-objective problem with an additional objective, the set size, and uses a simple evolutionary algorithm to optimise it. Such a Pareto optimisation subset selection can achieve the same general approximation guarantee as the greedy algorithm, but has better ability to avoid local optima.

5.7 Preference Incorporation from the DM

The ultimate goal of multiobjective optimisation is for the DM to select a single solution which most stands for their preferences. When the DM's preferences are unavailable, a set of well-converged, well-distributed solutions are favoured as they have a great probability of containing a solution picked by the DM. When the DM's preferences are available, it is certainly desirable to incorporate these preferences into the evaluation of solution sets.

Preference information of the DM can be manifold. When the preferences are clearly articulated, e.g., the DM preferring the knee of the problem or there being a hierarchy relation among objectives, quality evaluation of solution sets can be easily conducted on the basis of these preferences (see Li et al. (2018a)). When the articulation of the DM's preferences is general/rough, QIs need to be designed carefully to incorporate these preferences. Common ways of preference articulation in this regard are aspiration levels and reference weights. The aspiration levels are the levels on some/all objectives satisfied by the DM, and they usually result in a reference point in the objective space. To incorporate such information, indicators are designed to give the solutions that dominate or are close to the reference point (depending on the position of the point against the Pareto front) better evaluation results, such as those in Li and Deb (2016); Mohammadi et al. (2013); Yu et al. (2015). The weight articulation is to model the DM's preference by controlling the distribution of the weights in the space. This approach is also precisely the motivation behind some QIs, such as the R family Hansen and Jaszkiewicz (1998) and IPF Carlyle et al. (2003), as stated previously. Articulating the preferences by a set of weights seems promising, given the following advantages. First, it can be used to reflect the importance of different objectives and to control the distribution and range of solutions in the set. Second, it can easily be incorporated into many QIs (e.g., HV Zitzler et al. (2007) and $R2$ Wagner et al. (2013)). Finally, it works with/without the availability of preference information. Other types of preference articulation include the utility function and trade-off information. The former is

already incorporated in several indicators (see the R family and IPF), whereas the latter has rarely been considered in the literature.

5.8 Comprehensibility from User Perspective

Another issue of QIs is the comprehensibility of the evaluation outcome to the user. Based upon one or several scalar values returned by a QI evaluating solution sets, it could be difficult for the user to understand the quality of the solution sets, especially for those having no expertise in multiobjective optimisation. For example, given two solution sets $\mathbf{A}$ and $\mathbf{B}$, a quality indicator I evaluates $\mathbf{A}$ better than $\mathbf{B}$, say $I(\mathbf{A}) = 0.3 < I(\mathbf{B}) = 0.6$. Apparently, we cannot see much from this evaluation result. We do not know whether the DM will certainly choose the solution in $\mathbf{A}$ other than in $\mathbf{B}$ or not; or if the latter is the case, whether the chance of $\mathbf{A}$ being picked is twice as high as that of $\mathbf{B}$. As such, it is desirable to provide more meaningful interpretation of set evaluation, like what is the probability of the DM in favour of $\mathbf{A}$ (when the preference information of the DM is unavailable). Applying such an indicator to the above example, when $I(\mathbf{A}) = 0.3$ and $I(\mathbf{B}) = 0.6$, we know that the probability of $\mathbf{A}$ being picked is half as great as that of $\mathbf{B}$. If $I(\mathbf{A}) = 0.0$ and $I(\mathbf{B}) = 1.0$, we know that $\mathbf{A}$ is dominated by $\mathbf{B}$.

Table 5: A summary of representative quality indicators and their usage note/caveats and applicable situations.

Indicator	Quality aspect	Usage note/caveats	Applicable situation
$\mathcal{C}$	convergence	removing duplicate solutions before the calculation	more suitable for combinatorial problems where the Pareto front size is relatively small
GD	convergence	normalisation needed; replacing quadratic mean with arithmetic mean in the calculation	when the compared sets are nondominated to each other (i.e., no $\lhd$ relation between the sets, cf. Section 2.1), and the Pareto front range can be estimated properly (e.g., no DRS points)
GD$^+$	convergence	normalisation needed; replacing quadratic mean with arithmetic mean in the calculation	various situations
MS	spread	normalisation needed; removing the solutions that are dominated by some solution in the other sets before the calculation; this indicator only evaluates a set's extensity	when the compared sets are nondominated to each other
SP	uniformity	normalisation needed	when the compared sets are nondominated to each other
ER	cardinality		when the compared sets are nondominated to each other, and the sets have the same (or very similar) size
DCI	spread and uniformity	properly setting the grid division parameter	more accurate for the comparison between two solution sets
ϵ-indicator	comprehensive quality	normalisation needed; differently-performed sets may have the same/similar evaluation results	various situations, particularly for real-world problems
$R2$	comprehensive quality	normalisation needed; setting the reference point slightly better than the ideal point	various bi-objective situations
IGD	comprehensive quality	normalisation needed	when the compared sets are nondominated to each other, and a Pareto front representation with densely, uniformly distributed points is available
IGD$^+$	comprehensive quality	normalisation needed	when a Pareto front representation with densely and uniformly distributed points is available
HV	comprehensive quality	properly setting the reference point; exponentially increasing computational cost in objective dimensionality	when the objective dimensionality is not very high, and the problem's nadir point can be well estimated (e.g., in most benchmarking functions)

6 Conclusions

Problems to which multiobjective optimisers have been applied are ubiquitous in real life, and this suggests the need of careful analysis and fair evaluation of the optimisers' outcome. In this paper, we have carried out a comprehensive overview of quality evaluation in multiobjective optimisation. We have categorised 100 quality indicators, detailed several representative ones, and discussed some properties that quality indicators possess or are desirable to have. This all implies a conclusion that there is no ideal quality indicator to evaluate solution sets and different indicators fit in different situations. For instance, HV, thanks to its sensitiveness to dominance improvement, could be a good option when the problem's nadir point can be well estimated, e.g., in most benchmarking scenarios. $R2$ works for various bi-objective scenarios where we advise to use a reference point slightly better than the ideal point. The ϵ-indicator is an option for real world scenarios, as it is unaffected by the distribution and location of the reference set. GD and IGD needs to be replaced by GD^+ and IGD^+, respectively. The latter, in contrast to HV, can be used in benchmarking scenarios with the high objective dimensionality and/or the big set size provided that a densely and uniformly distributed reference set is available. Table 5 summarises the usage note/caveats and applicable situations of several quality indicators.

Finally, we have suggested several future research directions, in which linking quality evaluation with the DM's preferences needs much attention. In this regard, there are two paths forward. One is to design/modify quality indicators such that certain articulations of the DM's preferences are easily incorporated, such as in the R family Hansen and Jaszkiewicz (1998), IPF Carlyle et al. (2003), and HV Zitzler et al. (2007). The other is to consider to design/adapt quality indicators according to particular problem scenarios where the DM's preferences are given explicitly or implicitly, as in Li et al. (2018a).

7 Acknowledgement

This work was supported by National Key R&D Program of China (Grant No. 2017YFC0804003), EPSRC (Grant Nos. EP/J017515/1 and EP/P005578/1), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No. 2017ZT07X386), Shenzhen Peacock Plan (Grant No. KQTD2016112514355531), the Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. ZDSYS201703031748284) and the Program for University Key Laboratory of Guangdong Province (Grant No. 2017KSYS008).

References

S. F. Adra and P. J. Fleming. 2011. Diversity management in evolutionary many-objective optimization. *IEEE Transactions on Evolutionary Computation* 15, 2 (2011), 183–195.

I. Alaya, C. Solnon, and K. Ghedira. 2007. Ant colony optimization for multi-objective optimization problems. In *the 19th IEEE International Conference on Tools with Artificial Intelligence*, Vol. 1. 450–457.

M. Asafuddoula, T. Ray, and H. K. Singh. 2015. Characterizing Pareto front approximations in many-objective optimization. In *Proc. of the 2015 on Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 607–614.

A. Auger, J. Bader, and D. Brockhoff. 2010. Theoretically investigating optimal μ-distributions for the hypervolume indicator: first results for three objectives. In *Parallel Problem Solving from Nature (PPSN)*. 586–596.

A. Auger, J. Bader, D. Brockhoff, and E. Zitzler. 2009a. Investigating and exploiting the bias of the weighted hypervolume to articulate user preferences. In *Genetic and Evolutionary Computation Conference (GECCO)*. 563–570.

A. Auger, J. Bader, D. Brockhoff, and E. Zitzler. 2009b. Theory of the hypervolume indicator: Optimal μ-Distributions and the choice of the reference point. In *Proceedings of the 10th ACM SIGEVO workshop on Foundations of Genetic Algorithms (FOGA)*. 87–102.

J. Bader and E. Zitzler. 2011. HypE: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary Computation* 19, 1 (2011), 45–76.

S. Bandyopadhyay, S. K. Pal, and B. Aruna. 2004. Multiobjective GAs, quantitative indices, and pattern classification. *IEEE Transactions on Systems Man and Cybernetics, Part B (Cybernetics)* 34, 5 (2004), 2088–2099.

M. Basseur and E. K. Burke. 2007. Indicator-based multi-objective local search. In *IEEE Congress on Evolutionary Computation*. 3100–3107.

A. Berry and P. Vamplew. 2005. The combative accretion model–multiobjective optimisation without explicit Pareto ranking. In *Evolutionary Multi-Criterion Optimization*. Springer, 77–91.

N. Beume. 2009. S-metric calculation by considering dominated hypervolume as Klee's measure problem. *Evolutionary Computation* 17, 4 (Nov. 2009), 477–492.

N. Beume, C. M. Fonseca, M. Lopez-Ibanez, L. Paquete, and J. Vahrenhold. 2009. On the Complexity of Computing the Hypervolume Indicator. *IEEE Transactions on Evolutionary Computation* 13, 5 (2009), 1075–1082.

N. Beume, B. Naujoks, and Emmerich. 2007. SMS-EMOA: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research* 181, 3 (2007), 1653–1669.

P. AN Bosman and D. Thierens. 2003. The balance between proximity and diversity in multiobjective evolutionary algorithms. *IEEE Transactions on Evolutionary Computation* 7, 2 (2003), 174–188.

P. AN Bosman and D. Thierens. 2005. The naive MIDEA: A baseline multi-objective EA. In *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 428–442.

B. Bozkurt, J. W. Fowler, E. S. Gel, B. Kim, M. Köksalan, and J. Wallenius. 2010. Quantitative comparison of approximate solution sets for multicriteria optimization problems with weighted Tchebycheff preference function. *Operations Research* 58, 3 (2010), 650–659.

K. Bringmann. 2013. Bringing order to special cases of Klee's measure problem. In *International Symposium on Mathematical Foundations of Computer Science*. Springer, 207–218.

K. Bringmann, S. Cabello, and M. Emmerich. 2017. Maximum volume subset selection for anchored boxes. In *33rd International Symposium on Computational Geometry (SoCG 2017)*, Vol. 77. 22:1–22:15.

K. Bringmann and T. Friedrich. 2010a. Approximating the volume of unions and intersections of high-dimensional geometric objects. *Computational Geometry: Theory and Applications* 43, 6-7 (2010), 601–610.

K. Bringmann and T. Friedrich. 2010b. An efficient algorithm for computing hypervolume contributions. *Evolutionary Computation* 18, 3 (2010), 383–402.

K. Bringmann and T Friedrich. 2010c. The maximum hypervolume set yields near-optimal approximation. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO)*. ACM press, 511–518.

K. Bringmann and T. Friedrich. 2012. Approximating the least hypervolume contributor: NP-hard in general, but fast in practice. *Theoretical Computer Science* 425 (2012), 104–116.

K. Bringmann and T. Friedrich. 2013. Approximation quality of the hypervolume indicator. *Artificial Intelligence* 195 (2013), 265–290.

K. Bringmann, T. Friedrich, C. Igel, and Voß. 2013. Speeding up many-objective optimization by Monte Carlo approximations. *Artificial Intelligence* 204 (2013), 22–29.

K. Bringmann, T Friedrich, and P. Klitzke. 2014. Two-dimensional subset selection for hypervolume and epsilon-indicator. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO)*. ACM Press, 589–596.

D. Brockhoff, T. Wagner, and H. Trautmann. 2012. On the properties of the R2 indicator. In *Proceedings of Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 465–472.

D. Brockhoff, T. Wagner, and H. Trautmann. 2015. R2 indicator-based multiobjective search. *Evolutionary Computation* 23, 3 (2015), 369–395.

T. Brunsch and H. Röglin. 2015. Improved smoothed analysis of multiobjective optimization. *Journal of the ACM* 62, 1 (2015).

L. T. Bui, S. Wesolkowski, A. Bender, H. A. Abbass, and M. Barlow. 2009. A dominance-based stability measure for multi-objective evolutionary algorithms. In *IEEE Congress on Evolutionary Computation*. IEEE, 749–756.

X. Cai, H. Sun, and Z. Fan. 2018. A diversity indicator based on reference vectors for many-objective optimization. *Information Sciences* 430 (2018), 467–486.

Y. Cao, B. J. Smucker, and T. J. Robinson. 2015. On using the hypervolume indicator to compare Pareto fronts: Applications to multi-criteria optimal experimental design. *Journal of Statistical Planning and Inference* 160 (2015), 60–74.

W. M. Carlyle, J. W. Fowler, E. S. Gel, and B. Kim. 2003. Quantitative comparison of approximate solution sets for bi-criteria optimization problems. *Decision Sciences* 34, 1 (2003), 63–82.

T. M. Chan. 2013. Klee's measure problem made easy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS),*. IEEE, 410–419.

C. A. C. Coello. 2000. An updated survey of GA-based multiobjective optimization techniques. *ACM Comput. Surv.* 32, 2 (2000), 109–143.

C. A. C. Coello, G. B. Lamont, and D. A. Van Veldhuizen. 2007. *Evolutionary algorithms for solving multi-objective problems*. Vol. 5. New York: Springer.

C. A. C. Coello, G. T. Pulido, and M. S. Lechuga. 2004. Handling multiple objectives with particle swarm optimization. *IEEE Transactions on Evolutionary Computation* 8, 3 (2004), 256–279.

C. A. C. Coello and M. R. Sierra. 2004. A study of the parallelization of a coevolutionary multi-objective evolutionary algorithm. In *Proceedings of the Mexican International Conference on Artificial Intelligence (MICAI)*. 688–697.

Y. Collette and P. Siarry. 2005. Three new metrics to measure the convergence of metaheuristics towards the Pareto frontier and the aesthetic of a set of solutions in biobjective optimization. *Computers & Operations Research* 32, 4 (2005), 773–792.

P. Czyzak and A. Jaszkiewicz. 1998. Pareto simulated annealing – a metaheuristic technique for multiple-objective combinatorial optimization. *Journal of Multi-Criteria Decision Analysis* 7, 1 (1998), 34–47.

D. Datta and J. R. Figueira. 2012. Some convergence-based M-ary cardinal metrics for comparing performances of multi-objective optimizers. *Computers & Operations Research* 39, 7 (2012), 1754–1762.

P. De, J. B. Ghosh, and C. E. Wells. 1992. Heuristic estimation of the efficient frontier for a bi-criteria scheduling problem. *Decision Sciences* 23, 3 (1992), 596–609.

K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. 2000. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In *Parallel Problem Solving from Nature PPSN VI*. Springer Berlin / Heidelberg, 849–858.

K. Deb and S. Jain. 2002. *Running performance metrics for evolutionary multi-objective optimization*. Technical Report 2002004. KanGAL, Indian Institute of Technology.

K. Deb, M. Mohan, and S. Mishra. 2005. Evaluating the ϵ-domination based multi-objective evolutionary algorithm for a quick computation of Pareto-optimal solutions. *Evolutionary Computation* 13, 4 (Dec. 2005), 501–525.

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197.

A. Diaz-Manriquez, G. Toscano-Pulido, C. A. C. Coello, and R. Landa-Becerra. 2013. A ranking method based on the R2 indicator for many-objective optimization. In *2013 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1523–1530.

E. Dilettoso, S. A. Rizzo, and N. Salerno. 2017. A weakly Pareto compliant quality indicator. *Mathematical and Computational Applications* 22, 1 (2017), 25.

M. Ehrgott. 2006. *Multicriteria optimization*. Springer Science & Business Media.

A. E. Eiben and J. Smith. 2015. From evolutionary computation to the evolution of things. *Nature* 521, 7553 (2015), 476–482.

M. Emmerich, N. Beume, and B. Naujoks. 2005. An EMO algorithm using the hypervolume measure as selection criterion. In *Evolutionary Multi-Criterion Optimization*. Springer, 62–76.

M. Emmerich and A. Deutz. 2014. Time complexity and zeros of the hypervolume indicator gradient field.. In *In EVOLVE-A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation III*. Springer International Publishing, 169–193.

M. Emmerich, A. Deutz, and N. Beume. 2007. Gradient-based/evolutionary relay hybrid for computing Pareto front approximations maximizing the S-metric. In *Hybrid Metaheuristics*, T. et al. Bartz-Beielstein (Ed.). Springer, 140–156.

M. Emmerich, A. Deutz, and I. Yevseyeva. 2015. A Bayesian approach to portfolio selection in multicriteria group decision making. *Procedia Computer Science* 64 (2015), 993–1000.

M. Emmerich and C. M. Fonseca. 2011. Computing hypervolume contributions in low dimensions: asymptotically optimal algorithm and complexity results. In *Evolutionary Multi-Criterion Optimization (EMO)*. 121–135.

H. Eskandari, C. D. Geiger, and G. B. Lamont. 2007. FastPGA: A dynamic population sizing approach for solving expensive multiobjective optimization problems. In *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 141–155.

A. Farhang-Mehr and S. Azarm. 2003a. An information-theoretic metric for assessing multi-objective optimization solution set quality. *Transactions of the ASME, Journal of Mechanical Design* 125, 4 (2003), 655–663.

A. Farhang-Mehr and S. Azarm. 2003b. Minimal sets of quality metrics. In *Proceeding of the Second International Conference on Evolutionary Multi-Criterion Optimization (EMO)*. 405–417.

S. L. Faulkenberg and M. M. Wiecek. 2010. On the quality of discrete representations in multiple objective programming. *Optimization and Engineering* 11, 3 (2010), 423–440.

J. E. Fieldsend, R. M. Everson, and S. Singh. 2003. Using unconstrained elite archives for multiobjective optimization. *IEEE Transactions on Evolutionary Computation* 7, 3 (2003), 305–323.

M. Fleischer. 2003. The measure of Pareto optima applications to multi-objective metaheuristics. In *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 519–533.

C. M. Fonseca and P. J. Fleming. 1995. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation* 3, 1 (1995), 1–16.

C. M. Fonseca and P. J. Fleming. 1996. On the performance assessment and comparison of stochastic multiobjective optimizers. In *Proceedings of the International Conference on Parallel Problem Solving from Nature (PPSN)*. Vol. 1141. 584–593.

C. M. Fonseca, L. Paquete, and M. Lopez-Ibanez. 2006. An improved dimension-sweep algorithm for the hypervolume indicator. In *Proc. IEEE Congress Evolutionary Computation CEC 2006*. 1157–1163.

J. W. Fowler, B. Kim, W. M. Carlyle, E. S. Gel, and S. Horng. 2005. Evaluating solution sets of a posteriori solution techniques for bi-criteria combinatorial optimization problems. *Journal of Scheduling* 8, 1 (2005), 75–96.

T. Friedrich, C. Horoba, and F. Neumann. 2009. Multiplicative approximations and the hypervolume indicator. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO)*. 571–578.

C. K. Goh and K. C. Tan. 2007. An investigation on noisy environments in evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation* 11, 3 (2007), 354–381.

C. K. Goh and K. C. Tan. 2009. Evolutionary multi-objective optimization in uncertain environments. *Issues and Algorithms, Studies in Computational Intelligence* 186 (2009), 5–18.

D. E. Goldberg. 1989. *Genetic algorithms in search, optimization, and machine learning*. Addison-wesley.

A. P. Guerreiro and C. M. Fonseca. 2018. Computing and updating hypervolume contributions in up to four dimensions. *IEEE Transactions on Evolutionary Computation* 22, 3 (2018), 449–463.

A. P. Guerreiro, C. M. Fonseca, and M. Emmerich. 2012. A fast dimension-sweep algorithm for the hypervolume indicator in four dimensions. In *CCCG*. 77–82.

D. Hadka and P. Reed. 2012. Diagnostic assessment of search controls and failure modes in many-objective evolutionary optimization. *Evolutionary Computation* 20, 3 (2012), 423–452.

M. P. Hansen. 1997. Tabu search for multiobjective optimization: MOTS. In *Proceedings of the 13th international conference on multiple criteria decision making*. 574–586.

M. P. Hansen and A. Jaszkiewicz. 1998. *Evaluating the quality of approximations to the nondominated set*. Imm-rep-1998-7. Institute of Mathematical Modeling, Technical University of Denmark.

D. P. Hardin and E. B. Saff. 2004. Discretizing manifolds via minimum energy points. *Notices of the AMS* 51, 10 (2004), 1186–1194.

Z. He and G. G. Yen. 2016. Visualization and performance metric in many-objective optimization. *IEEE Transactions on Evolutionary Computation* 20, 3 (2016), 386–402.

M. Helbig and A. P. Engelbrecht. 2014. Benchmarks for dynamic multi-objective optimisation algorithms. *ACM Comput. Surv.* 46, 3, Article 37 (Jan. 2014), 39 pages.

S. Helbig and D. Pateva. 1994. On several concepts for ϵ-efficiency. *Operations Research Spektrum* 16, 3 (1994), 179–186.

G. R. Hernández and C. C. A. Coello. 2013. MOMBI: A new metaheuristic for many-objective optimization based on the R2 indicator. In *2013 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2488–2495.

G. R. Hernández and C. C. A. Coello. 2015. Improved metaheuristic based on the R2 indicator for many-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 679–686.

R. M. Hierons, M. Li, X. Liu, S. Segura, and W. Zheng. 2016. SIP: optimal product selection from feature models using many-objective evolutionary optimization. *ACM Transactions on Software Engineering and Methodology* 25, 2 (2016), 17.

T. Hiroyasu, M. Miki, and S. Watanabe. 2000. The new model of parallel genetic algorithm in multi-objective optimization problems-divided range multi-objective genetic algorithm. In *Proceedings of the 2000 Congress on Evolutionary Computation*, Vol. 1. IEEE, 333–340.

I. Hupkens and M. Emmerich. 2013. Logarithmic-time updates in SMS-EMOA and hypervolume-based archiving. In *EVOLVE-A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation IV*. Springer International Publishing, 155–169.

A. Ibrahim, S. Rahnamayan, M. V. Martin, and K. Deb. 2017. 3D-RadVis Antenna: Visualization and Performance Measure for Many-objective Optimization. *Swarm and Evolutionary Computation* (2017).

C. Igel, N. Hansen, and S. Roth. 2007. Covariance matrix adaptation for multi-objective optimization. *Evolutionary Computation* 15, 1 (March 2007), 1–28.

K. Ikeda, H. Kita, and S. Kobayashi. 2001. Failure of Pareto-based MOEAs: does non-dominated really mean near to optimal?. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, Vol. 2. 957–962.

A. Inselberg and B. Dimsdale. 1991. Parallel coordinates. In *Human-Machine Interactive Systems*. Springer, 199–233.

H. Ishibuchi, R. Imada, Y. Setoguchi, and Y. Nojima. 2017. Reference point specification in hypervolume calculation for fair comparison and efficient search. In *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 585–592.

H. Ishibuchi, R. Imada, Y. Setoguchi, and Y. Nojima. 2018a. How to specify a reference point in hypervolume calculation for fair performance comparison. *Evolutionary Computation* 26, 3 (2018), 411–440.

H. Ishibuchi, R. Imada, Y. Setoguchi, and Y. Nojima. 2018b. Reference point specification in inverted generational distance for triangular linear Pareto front. *IEEE Transactions on Evolutionary Computation* (2018).

H. Ishibuchi, H. Masuda, and Y. Nojima. 2015. A study on performance evaluation ability of a modified inverted generational distance indicator. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*. 695–702.

H. Ishibuchi, H. Masuda, Y. Tanigaki, and Y. Nojima. 2014. Difficulties in specifying reference points to calculate the inverted generational distance for many-objective optimization problems. In *IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making (MCDM)*. 170–177.

H. Ishibuchi, H. Masuda, Y. Tanigaki, and Y. Nojima. 2015. Modified distance calculation in generational distance and inverted generational distance. In *Proceedings of the International Conference on Evolutionary Multi-Criterion Optimization (EMO)*. 110–125.

H. Ishibuchi and T. Murata. 1998. A multi-objective genetic local search algorithm and its application to flowshop scheduling. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews* 28, 3 (1998), 392–403.

H. Ishibuchi and Y. Shibata. 2004. Mating scheme for controlling the diversity-convergence balance for multiobjective optimization. In *Genetic and Evolutionary Computation Conference (GECCO)*. 1259–1271.

H. Ishibuchi, N. Tsukamoto, and Y. Nojima. 2008. Evolutionary many-objective optimization: A short review. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*. 2419–2426.

H. Ishibuchi, N. Tsukamoto, Y. Sakane, and Y. Nojima. 2010. Indicator-based evolutionary algorithm with hypervolume approximation by achievement scalarizing functions. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 527–534.

H. Ishibuchi, T. Yoshida, and T. Murata. 2003. Balance between genetic search and local search in memetic algorithms for multiobjective permutation flowshop scheduling. *IEEE Transactions on Evolutionary Computation* 7, 2 (2003), 204–223.

A. L. Jaimes and C. C. A. Coello. 2009. Study of preference relations in many-objective optimization. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO)*. 611–618.

A. Jaszkiewicz. 2018. Improved quick hypervolume algorithm. *Computers & Operations Research* 90 (2018), 72–83.

S. Jiang, Y. S. Ong, J. Zhang, and L. Feng. 2014. Consistencies and contradictions of performance metrics in multiobjective optimization. *IEEE Transactions on Cybernetics* 44, 12 (2014), 2391–2404.

S. Jiang, S. Yang, and M. Li. 2016. On the use of hypervolume for diversity measurement of Pareto front approximations. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. 1–8.

S. Jiang, J. Zhang, Y. S. Ong, A. N. Zhang, and P. S. Tan. 2015. A simple and fast hypervolume indicator-based multiobjective evolutionary algorithm. *IEEE Transactions on Cybernetics* 45, 10 (2015), 2202–2213.

D. F. Jones, S. K. Mirrazavi, and M. Tamiz. 2002. Multi-objective meta-heuristics: An overview of the current state-of-the-art. *European Journal of Operational Research* 137, 1 (2002), 1 – 9.

H. Kaji and H. Kita. 2007. Individual evaluation scheduling for experiment-based evolutionary multi-objective optimization. In *Evolutionary Multi-Criterion Optimization*. Springer, 645–659.

E. Kasanen, R. Östermark, and M. Zeleny. 1991. Gestalt system of holistic graphics: new management support view of MCDM. *Computers & Operations Research* 18, 2 (1991), 233–239.

B. Kim, E. S. Gel, J. W. Fowler, W. M. Carlyle, and J. Wallenius. 2006. Evaluation of nondominated solution sets for k-objective optimization problems: An exact method and approximations. *European Journal of Operational Research* 173, 2 (2006), 565–582.

J. D. Knowles. 2002. *Local-search and hybrid evolutionary algorithms for Pareto optimization*. Ph.D. Dissertation. University of Reading UK.

J. D. Knowles and D. W. Corne. 1999. The Pareto archived evolution strategy: a new baseline algorithm for Pareto multiobjective optimisation. In *Proc. Congress Evolutionary Computation CEC 99*, Vol. 1.

J. D. Knowles and D. W. Corne. 2002. On metrics for comparing nondominated sets. In *Proceeding of the Congress Evolutionary Computation (CEC)*, Vol. 1. 711–716.

J. D. Knowles and D. W. Corne. 2003. Properties of an adaptive archiving algorithm for storing nondominated vectors. *IEEE Transactions on Evolutionary Computation* 7, 2 (2003), 100–116.

J. D. Knowles, D. W. Corne, and M. Fleischer. 2003. Bounded archiving using the Lebesgue measure. In *Congress on Evolutionary Computation (CEC)*. 2490–2497.

J. D. Knowles, L. Thiele, and E. Zitzler. 2006. *A tutorial on the performance assessment of stochastic multiobjective optimizers*. Technical Report No. 214. Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Switzerland.

J. Kollat and P. Reed. 2005. The value of online adaptive search: a performance comparison of NSGAII, ε-NSGAII and εMOEA. In *Evolutionary Multi-Criterion Optimization*. Springer, 386–398.

T. Kuhn, C. M. Fonseca, L. Paquete, S. Ruzika, M. M. Duarte, and J. R. Figueira. 2016. Hypervolume subset selection in two dimensions: Formulations and algorithms. *Evolutionary Computation* 24, 3 (2016), 411–425.

R. Lacour, K. Klamroth, and C. M. Fonseca. 2017. A box decomposition algorithm to compute the hypervolume indicator. *Computers & Operations Research* 79 (2017), 347–360.

M. Laumanns and R. Zenklusen. 2011. Stochastic convergence of random search methods to fixed size Pareto front approximations. *European Journal of Operational Research* 213, 2 (2011), 414–421.

Y. W. Leung and Y. Wang. 2003. U-measure: a quality measure for multiobjective programming. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 33, 3 (2003), 337–343.

B. Li, J. Li, K. Tang, and X. Yao. 2015. Many-objective evolutionary algorithms: A survey. *Comput. Surveys* 48, 1 (2015), 1–35.

B. Li, K. Tang, J. Li, and X. Yao. 2016. Stochastic ranking algorithm for many-objective optimization based on multiple indicators. *IEEE Transactions on Evolutionary Computation* 20, 6 (2016), 924–938.

K. Li and K. Deb. 2016. Performance assessment for preference-based evolutionary multi-objective optimization using reference points. *COIN Report* 1, 1 (2016), 1–23.

M. Li, T. Chen, and X. Yao. 2018a. A critical review of "A practical guide to select quality indicators for assessing Pareto-based search algorithms in search-based software engineering": Essay on quality indicator selection for SBSE. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results Track*.

M. Li, C. Grosan, S. Yang, X. Liu, and X. Yao. 2018b. Multi-line distance minimization: A visualized many-objective test problem suite. *IEEE Transactions on Evolutionary Computation* 22, 1 (2018), 61–78.

M. Li, L. Hu, and X. Yao. 2019. On comparing two sets of multi-objective solution vectors. *arXiv preprint arXiv:1702.00477* (2019).

M. Li, S. Yang, and X. Liu. 2014a. Diversity comparison of Pareto front approximations in many-objective optimization. *IEEE Transactions on Cybernetics* 44, 12 (2014), 2568–2584.

M. Li, S. Yang, and X. Liu. 2014b. Shift-based density estimation for Pareto-based algorithms in many-objective optimization. *IEEE Transactions on Evolutionary Computation* 18, 3 (2014), 348–365.

M. Li, S. Yang, and X. Liu. 2015a. A performance comparison indicator for Pareto front approximations in many-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*. 703–710.

M. Li, S. Yang, and X. Liu. 2015b. Bi-goal evolution for many-objective optimization problems. *Artificial Intelligence* 228 (2015), 45–65.

M. Li, S. Yang, J. Zheng, and X. Liu. 2014. ETEA: A Euclidean minimum spanning tree-based evolutionary algorithm for multiobjective optimization. *Evol. Comput.* 22, 2 (2014), 189–230.

M. Li and X. Yao. 2019. An empirical investigation of the optimality and monotonicity properties of multiobjective archiving methods. In *Proc. Evolutionary Multi-Criterion Optimization (EMO), in press*.

M. Li, L. Zhen, and X. Yao. 2017. How to read many-objective solution sets in parallel coordinates. *IEEE Computational Intelligence Magazine* 12, 4 (2017), 88–97.

M. Li and J. Zheng. 2009. Spread assessment for evolutionary multi-objective optimization. In *Proc. Evolutionary Multi-Criterion Optimization (EMO)*. Nantes, France, 216–230.

M. Li, J. Zheng, and G. Xiao. 2008. Uniformity assessment for evolutionary multi-objective optimization. In *Proc. IEEE Congress Evolutionary Computation (CEC 2008)*. 625–632.

Xu-yong Li, Jin-hua Zheng, and Juan Xue. 2005. A diversity metric for multi-objective evolutionary algorithms. In *International Conference on Natural Computation*. 68–73.

A. Liefooghe and B. Derbel. 2016. A correlation analysis of set quality indicator values in multiobjective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 581–588.

G. Lizárraga-Lizárraga, A. Hernández-Aguirre, and S. Botello-Rionda. 2008a. G-Metric: an M-ary quality indicator for the evaluation of non-dominated sets. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. 665–672.

G. Lizárraga-Lizárraga, A. Hernández-Aguirre, and S. Botello-Rionda. 2008b. On the possibility to create a compatible–complete unary comparison method for evolutionary multiobjective algorithms. In *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*. 759–760.

G. Lizárraga-Lizárraga, A. Hernández-Aguirre, and S. Botello-Rionda. 2008c. Some Demonstrations about the Cardinality of Important Sets of Non-dominated Sets. In *Mexican International Conference on Artificial Intelligence*. 440–450.

E. M. Lopez and C. A. C. Coello. 2016. IGD+-EMOA: A multi-objective evolutionary algorithm based on IGD+. In *IEEE Congress on Evolutionary Computation*. IEEE, 999–1006.

A. V. Lotov, V. A. Bushenkov, and G. K. Kamenev. 2013. *Interactive decision maps: Approximation and visualization of Pareto frontier*. Vol. 89. Springer Science & Business Media.

A. V. Lotov, G. K. Kamenev, and V. E. Berezkin. 2002. Approximation and visualization of the Pareto frontier for nonconvex multicriteria problems. In *Doklady Mathematics*, Vol. 66. MAIK Nauka/Interperiodica, 260–262.

L. Mandow and J. L. P. De La Cruz. 2010. Multiobjective A* search with consistent heuristics. *Journal of the ACM* 57, 5 (2010), 27.

H. Meng, X. Zhang, and S. Liu. 2005. New quality measures for multiobjective programming. *Advances in Natural Computation* (2005), 431–431.

A. Messac and C. A. Mattson. 2004. Normal constraint method with guarantee of even representation of complete Pareto frontier. *AIAA journal* 42, 10 (2004), 2101–2111.

H. Meunier, E. G. Talbi, and P. Reininger. 2000. A multiobjective genetic algorithm for radio network optimization. In *Proceedings of the 2000 Congress on Evolutionary Computation*, Vol. 1. 317–324.

K. Miettinen. 1999. *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, Boston.

K. Miettinen. 2014. Survey of methods to visualize alternatives in multiple criteria decision making problems. *OR Spectrum* 36 (2014), 3–37.

A. Mohammadi, M. N. Omidvar, and X. Li. 2013. A new performance metric for user-preference based multi-objective evolutionary algorithms. In *IEEE Congress on Evolutionary Computation*. IEEE, 2825–2832.

S. Mostaghim and J. Teich. 2005. A new approach on many objective diversity measurement. In *Practical Approaches to Multi-Objective Optimization*.

A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello. 2014a. A survey of multiobjective evolutionary algorithms for data mining: Part I. *IEEE Transactions on Evolutionary Computation* 18, 1 (2014), 4–19.

A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello. 2014b. Survey of multiobjective evolutionary algorithms for data mining: Part II. *IEEE Transactions on Evolutionary Computation* 18, 1 (Feb 2014), 20–35.

M. Nicolini. 2004. Evaluating performance of multi-objective genetic algorithms for water distribution system optimization. In *the 6th International Conference on Hydroinformatic*. World Scientific, 850–857.

T. Okabe, Y. Jin, and B. Sendhoff. 2003. A critical survey of performance indices for multi-objective optimisation. In *Proceeding of the Congress Evolutionary Computation (CEC)*, Vol. 2. 878–885.

T. Pamulapati, R. Mallipeddi, and P. N. Suganthan. 2018. I^+_{SDE} - An indicator for multi and many-objective optimization. *IEEE Transactions on Evolutionary Computation, in press* (2018).

C. H. Papadimitriou and M. Yannakakis. 2000. On the approximability of trade-offs and optimal access of web sources. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*. 86–92.

D. H. Phan and J. Suzuki. 2013. R2-IBEA: R2 indicator based evolutionary algorithm for multiobjective optimization. In *IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1836–1845.

D. H. Phan, J. Suzuki, and I. Hayashi. 2012. Leveraging indicator-based ensemble selection in evolutionary multiobjective optimization algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 497–504.

A. Ponsich, A. L. Jaimes, and C. A. C. Coello. 2013. A survey on multiobjective evolutionary algorithms for the solution of the portfolio optimization problem and other finance and economics applications. *IEEE Transactions on Evolutionary Computation* 17, 3 (2013), 321–344.

R. C. Purshouse and P. J. Fleming. 2007. On the evolutionary optimization of many conflicting objectives. *IEEE Transactions on Evolutionary Computation* 11, 6 (2007), 770–784.

C. Qian, G. Li, C. Feng, and K. Tang. 2018. Distributed Pareto optimization for subset selection.. In *International Joint Conference on Artificial Intelligence*. 1492–1498.

C. Qian, J. Shi, Y. Yu, K. Tang, and Z. Zhou. 2016. Parallel Pareto optimization for subset selection. In *International Joint Conference on Artificial Intelligence*. 1939–1945.

C. Qian, Y. Yu, and Z. Zhou. 2015. Subset selection by Pareto optimization. In *Advances in Neural Information Processing Systems*. 1774–1782.

M. Ravber, M. Mernik, and M. Črepinšek. 2017. The impact of quality indicators on the rating of multi-objective evolutionary algorithms. *Applied Soft Computing* 55 (2017), 265–275.

N. Riquelme, C. Von Lücken, and B. Baran. 2015. Performance metrics in multi-objective optimization. In *Latin American Computing Conference (CLEI)*. 1–11.

G. Rote, K. Buchin, K. Bringmann, S. Cabello, and M. Emmerich. 2016. Selecting k points that maximize the convex hull volume. In *The 19th Japan Conference on Discrete and Computational Geometry, Graphs, and Games*. 58–60.

G. Rudolph, O. Schütze, C. Grimme, C. Domínguez-Medina, and H. Trautmann. 2016. Optimal averaged Hausdorff archives for bi-objective problems: theoretical and numerical results. *Computational Optimization and Applications* 64, 2 (2016), 589–618.

G. Rudolph, O. Schütze, C. Grimme, and H. Trautmann. 2014. An aspiration set EMOA based on averaged Hausdorff distances. In *International Conference on Learning and Intelligent Optimization*. 153–156.

L. M. S. Russo and A. P. Francisco. 2014. Quick hypervolume. *IEEE Transactions on Evolutionary Computation* 18, 4 (2014), 481–502.

S. Sayın. 2000. Measuring the quality of discrete representations of efficient sets in multiple objective mathematical programming. *Mathematical Programming* 87, 3 (2000), 543–560.

J. R. Schott. 1995. *Fault tolerant design using single and multicriteria genetic algorithm optimization*. Master's thesis. Department of Aeronautics and Astronautics, Massachusetts Institute of Technology.

O. Schutze, X. Esquivel, A. Lara, and C. C. A. Coello. 2012. Using the averaged Hausdorff distance as a performance measure in evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation* 16, 4 (2012), 504–522.

M. R. Sierra and C. A. C. Coello. 2005. Improving PSO-based multi-objective optimization using crowding, mutation and ϵ-dominance. In *Evolutionary multi-criterion optimization*, Vol. 3410. 505–519.

N. Srinivas and K. Deb. 1994. Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation* 2, 3 (1994), 221–248.

Y. Sun, G. G. Yen, and Y. Zhang. 2018. IGD indicator-based evolutionary algorithm for many-objective optimization problems. *IEEE Transactions on Evolutionary Computation* published online (2018).

K. C. Tan, T. H. Lee, and E. F. Khor. 2002. Evolutionary algorithms for multi-objective optimization: Performance assessments and comparisons. *Artificial Intelligence Review* 17 (2002), 251–290. Issue 4.

L. Thiele. 2015. Indicator-based selection. In *Springer Handbook of Computational Intelligence*. Springer, 983–994.

L. Thiele, K. Miettinen, P. J. Korhonen, and J. Molina. 2009. A preference-based evolutionary algorithm for multi-objective optimization. *Evolutionary computation* 17, 3 (2009), 411–436.

Y. Tian, R. Cheng, X. Zhang, F. Cheng, and Y. Jin. 2017. An indicator based multi-objective evolutionary algorithm with reference point adaptation for better versatility. *IEEE Transactions on Evolutionary Computation* published online (2017).

Y. Tian, R. Cheng, X. Zhang, M. Li, and Y. Jin. 2019. Diversity assessment of multi-objective evolutionary algorithms: Performance metric and benchmark problems. *IEEE Computational Intelligence Magazine, submitted* (2019).

Y. Tian, X. Zhang, R. Cheng, and Y. Jin. 2016. A multi-objective evolutionary algorithm based on an enhanced inverted generational distance metric. In *IEEE Congress on Evolutionary Computation (CEC)*. 5222–5229.

H. Trautmann, T. Wagner, and D. Brockhoff. 2013. R2-EMOA: Focused multiobjective search using R2-indicator-based selection. In *International Conference on Learning and Intelligent Optimization*. 70–74.

A. Trivedi, D. Srinivasan, K. Sanyal, and A. Ghosh. 2017. A survey of multiobjective evolutionary algorithms based on decomposition. *IEEE Transactions on Evolutionary Computation* 21, 3 (2017), 440–462.

P. A. Tukey and J. W. Tukey. 1981. *Preparation; prechosen sequences of views*. Wiley, Chapter Interpreting multivariate data, 189–213.

T. Tusar and B. Filipic. 2015. Visualization of Pareto front approximations in evolutionary multiobjective optimization: A critical review and the prosection method. *IEEE Transactions on Evolutionary Computation* 19, 2 (2015), 225–245.

E. L. Ulungu, J. Teghem, P. H. Fortemps, and D. Tuyttens. 1999. MOSA method: a tool for solving multiobjective combinatorial optimization problems. *Journal of multicriteria decision analysis* 8, 4 (1999), 221.

D. A. Van Veldhuizen. 1999. *Multiobjective evolutionary algorithms: Classifications, analyses, and new innovations*. Ph.D. Dissertation. Department of Electrical and Computer Engineering, Graduate School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

D. A. Van Veldhuizen and G. B. Lamont. 1998. Evolutionary computation and convergence to a Pareto front. In *Late Breaking Papers at the Genetic Programming Conference*. 221–228.

D. Vaz, L. Paquete, C. M. Fonseca, K. Klamroth, and M. Stiglmayr. 2015. Representation of the non-dominated set in biobjective discrete optimization. *Computers & Operations Research* 63 (2015), 172–186.

D. Vaz, L. Paquete, and A. Ponte. 2013. A note on the epslion-indicator subset selection. *Theoretical Computer Science* 499 (2013), 113–116.

A. Viana and J. P. de Sousa. 2000. Using metaheuristics in multiobjective resource constrained project scheduling. *European Journal of Operational Research* 120, 2 (2000), 359–374.

C. A. R. Villalobos and C. C. A. Coello. 2012. A new multi-objective evolutionary algorithm based on a performance assessment indicator. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO)*. ACM, 505–512.

T. Wagner, H. Trautmann, and D. Brockhoff. 2013. *Preference articulation by means of the R2 indicator*. 81–95.

D. J. Walker, R. M. Everson, and J. E. Fieldsend. 2013. Visualising mutually non-dominating solution sets in many-objective optimisation. *IEEE Transactions on Evolutionary Computation* 17, 2 (2013), 165–184.

J. Wallenius, J. S. Dyer, P. C. Fishburn, R. E. Steuer, S. Zionts, and K. Deb. 2008. Multiple criteria decision making, multiattribute utility theory: Recent accomplishments and what lies ahead. *Management science* 54, 7 (2008), 1336–1349.

H. Wang, Y. Jin, and X. Yao. 2017. Diversity assessment in many-objective optimization. *IEEE transactions on cybernetics* 47, 6 (2017), 1510–1522.

R. Wang, S. Chen, L. Ma, S. Cheng, and Y. Shi. 2018. Multi-indicator bacterial foraging algorithm with Kriging model for many-objective optimization. In *International Conference on Swarm Intelligence*. 530–539.

S. Wang, S. Ali, T. Yue, Y. Li, and M. Liaaen. 2016. A practical guide to select quality indicators for assessing Pareto-based search algorithms in search-based software engineering. In *Proceedings of the 38th International Conference on Software Engineering (ICSE)*. 631–642.

S. Wessing and B. Naujoks. 2010. Sequential parameter optimization for multi-objective problems. In *IEEE Congress on Evolutionary Computation*. 1–8.

L. While, L. Bradstreet, and L. Barone. 2012. A fast way of calculating exact hypervolumes. *IEEE Transactions on Evolutionary Computation* 16, 1 (2012), 86–95.

L. While, P. Hingston, L. Barone, and S. Huband. 2006. A faster algorithm for calculating hypervolume. *IEEE Transactions on Evolutionary Computation* 10, 1 (2006), 29–38.

J. Wu and S. Azarm. 2001. Metrics for quality assessment of a multiobjective design optimization solution set. *Transactions of the ASME, Journal of Mechanical Design* 123 (2001), 18–25.

Y. Xiang, Y. Zhou, Z. Zheng, and M. Li. 2018. Configuring software product lines by combining many-objective optimization and SAT solvers. *ACM Transactions on Software Engineering and Methodology* 26, 4 (2018), 14.

S. Yang, M. Li, X. Liu, and J. Zheng. 2013. A grid-based evolutionary algorithm for many-objective optimization. *IEEE Transactions on Evolutionary Computation* 17, 5 (2013), 721–736.

G. G. Yen and Z. He. 2014. Performance metric ensemble for multiobjective evolutionary algorithms. *IEEE Transactions on Evolutionary Computation* 18, 1 (2014), 131–144.

I. Yevseyeva, A. P. Guerreiro, M. Emmerich, and C. M. Fonseca. 2014. A portfolio optimization approach to selection in multiobjective evolutionary algorithms. In *Parallel Problem Solving from Nature (PPSN)*. 672–681.

H. Yildiz and S. Suri. 2012. On Klee's measure problem for grounded boxes. In *Proceedings of the twenty-eighth annual symposium on Computational geometry*. ACM, 111–120.

G. Yu, J. Zheng, and X. Li. 2015. An improved performance metric for multiobjective evolutionary algorithms with user preferences. In *IEEE Congress on Evolutionary Computation*. 908–915.

M. Zeleny. 1973. Compromise programming. *Multiple criteria decision making* (1973).

A. Zhou, Y. Jin, Q. Zhang, B. Sendhoff, and E. Tsang. 2006. Combining model-based and genetics-based offspring generation for multi-objective optimization using a convergence criterion. In *IEEE Congress on Evolutionary Computation*. 892–899.

A. Zhou, B. Y. Qu, H. Li, S. Z. Zhao, P. N. Suganthan, and Q. Zhang. 2011. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation* 1, 1 (2011), 32–49.

E. Zitzler. 1999. *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*. Ph.D. Dissertation. Zurich, Switzerland: Swiss Federal Institute of Technology (ETH).

E. Zitzler, D. Brockhoff, and L. Thiele. 2007. The hypervolume indicator revisited: On the design of Pareto-compliant indicators via weighted integration. In *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 862–876.

E. Zitzler, K. Deb, and L. Thiele. 2000. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation* 8, 2 (2000), 173–195.

E. Zitzler, J. Knowles, and L. Thiele. 2008. Quality assessment of Pareto set approximations. In *Multiobjective Optimization*, J. Branke, Kalyanmoy Deb, Kaisa Miettinen, and Roman Slowinski (Eds.). Vol. 5252. Springer Berlin / Heidelberg, 373–404.

E. Zitzler and S. Künzli. 2004. Indicator-based selection in multiobjective search. In *Proceedings of the International Conference on Parallel Problem Solving from Nature (PPSN)*. 832–842.

E. Zitzler and L. Thiele. 1998. Multiobjective optimization using evolutionary algorithms - A comparative case study. In *Proceedings of the International Conference on Parallel Problem Solving from Nature (PPSN)*. 292–301.

E. Zitzler and L. Thiele. 1999. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation* 3, 4 (1999), 257–271.

E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. Da Fonseca. 2003. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation* 7, 2 (2003), 117–132.