Information Sciences 324 (2015) 270–285

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

On the estimation of Pareto fronts from the point of view of copula theory

Mickaël Binois [a,b,*], Didier Rullière [c], Olivier Roustant [b]

[a] Renault S.A.S., Guyancourt 78084, France
[b] Mines Saint-Étienne, UMR CNRS 6158, LIMOS, F-42023 Saint-Étienne, France
[c] Université de Lyon, Université Lyon 1, ISFA, Laboratoire SAF, EA2429, 50 avenue Tony Garnier, Lyon 69366, France

ARTICLE INFO

Article history:
Received 22 January 2015
Revised 2 May 2015
Accepted 22 June 2015
Available online 2 July 2015

Keywords:
Multi-objective optimization
Pareto front
Copulas
Archimedean copulas

ABSTRACT

Given a first set of observations from a design of experiments sampled randomly in the design space, the corresponding set of non-dominated points usually does not give a good approximation of the Pareto front. We propose here to study this problem from the point of view of multivariate analysis, introducing a probabilistic framework with the use of copulas. This approach enables the expression of level lines in the objective space, giving an estimation of the position of the Pareto front when the level tends to zero. In particular, when it is possible to use Archimedean copulas, analytical expressions for Pareto front estimators are available. Several case studies illustrate the interest of the approach, which can be used at the beginning of the optimization when sampling randomly in the design space.

1. Introduction

Multi-objective optimization (MOO) received a lot of attention recently, including in particular developments on scalarization [22], hybrid approaches [25], evolutionary optimization (see e.g. [7,10,52]) or surrogate-based optimization [48]. Since no solution usually minimizes every objective at once, the definition of a solution for a multi-objective optimization problem is generally defined as a compromise: a solution is said to be optimal in the Pareto sense if there exists no other solution which is better for every component. All the optimal points in the objective space form the Pareto front. As a result, optimizers provide a set of non-dominated points to approximate the Pareto front. Methods are then designed to seek some properties for these sets, such as uniformity and coverage.

Usually an optimization process starts with random sampling, either to generate an initial population or as a basis to construct a metamodel. The current Pareto front estimated from this first sample may be highly variable, especially when only a small number of function evaluations are available, corresponding to time-consuming functions. This is illustrated in Fig. 1 for the bi-objective problem ZDT1 [53], with five 50-points initial samples. However, the stochastic nature of sampling provides a probabilistic framework that can be exploited to quantify this variability and to give a better initial localization of the Pareto front. More precisely, if $\mathbf{X} = (X_1, \ldots, X_d)$ is a d-dimensional random vector representing the inputs, and $f_1, \ldots, f_m$ the objective functions, then the Pareto front should be connected to the extreme level lines of the distribution of $\mathbf{Y} = (f_1(\mathbf{X}), \ldots, f_m(\mathbf{X}))$. To investigate such connection is the aim of the paper.

* Corresponding author at: Renault S.A.S., 78084 Guyancourt, France. Tel.: +33 477420285.
E-mail addresses: binois@emse.fr (M. Binois), didier.rulliere@univ-lyon1.fr (D. Rullière), roustant@emse.fr (O. Roustant).

http://dx.doi.org/10.1016/j.ins.2015.06.037
0020-0255/

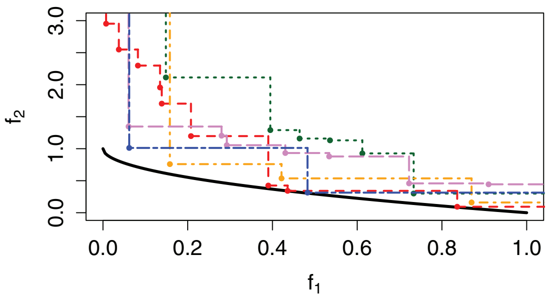

Fig. 1. Non-dominated points obtained with 5 different random samples (one color and type of line per sample) of 50 points for the bi-objective problem ZDT1. The true Pareto front is the black solid line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

In the mono-objective situation, a similar probabilistic connection is studied by [46] to estimate the value of the extremum. Considering a small sample of n observations $(y_1, \dots, y_n)$ of $\mathbf{Y}$, the minimum of $\mathbf{Y}$ is approximated using concepts from extreme order statistics. In multi-objective optimization, the connection seems to be new. Uncertainty quantification around the Pareto front has been recently considered by [4], using conditional simulations of Kriging metamodels and concepts from random sets theory. Whereas such approach is relevant in a sequential algorithm, it may be inappropriate in the initial stage that we consider here, due to a potentially large model error in metamodeling.

In this paper, we give a theoretical framework in which the Pareto front appears as a zero level line of the multivariate distribution $F_{\mathbf{Y}}$ of $\mathbf{Y}$. This problem is known in the probabilistic literature as support curve estimation (see e.g. [23,26,28]). However, the existing methods rely on assumptions, such as domain of attraction or polynomial rate of decrease, which can hardly be checked in an optimization context. As an alternative, we propose to take advantage of copulas [41] which are multivariate probability distributions with uniform marginals, allowing to consider separately the estimation of the marginals and the dependence structure. This allows estimating extreme level lines, without making specific assumptions about domain of attractions. Copulas have already been used in optimization, mainly in the variable space to estimate distribution in evolutionary algorithms, see e.g. [18,19,49], while here we focus on the objective space. We propose a first estimation of the Pareto front relying on the empirical copula. Then, we consider the case where the copula belongs to the class of Archimedean copulas, parameterized by a function. This assumption can be checked visually or statistically with specific tests of the literature. If relevant, a better localization of the Pareto front is found. Furthermore, a parametric expression of the approximated Pareto front is available.

The paper is structured as follows. Section 2 proposes alternative definitions of the Pareto front from the point of view of the cumulative distribution function, presents some background about copulas and describes the estimation procedure in the Archimedean case. Section 3 discusses the applicability of the model and more specifically the consequences of the Archimedean copula model. Section 4 illustrates in several configurations the application of the proposed approach to Pareto front localization. Section 5 concludes and describes possibilities for further improvements.

2. Methodology

The present section describes the interest of using a probabilistic framework in multi-objective optimization by establishing the link between both domains. Based on the resulting theorem, the expression of level lines of the multivariate cumulative distribution functions $F_{\mathbf{Y}}$ using copulas is described as well as a procedure for their estimation. Empirical and parametric models are discussed, with emphasis on Archimedean models.

2.1. Link between Pareto front and level curves

For a variety of methods ranging from evolutionary optimization [10] to surrogate-based methods [45], optimization starts with random sampling in the design space, with uniform sampling or with a random Latin Hypercube. In this case, it is possible to study the resulting observations in the objective space as a set of points. Specifically, assuming that the outputs can be considered as independent and identically distributed (i.i.d.) random variables, they enter the scope of multivariate analysis.

Let us start with definitions of Pareto dominance and Pareto front, in a minimization context. For two points $\mathbf{y} = (y_1, \dots, y_m)$ and $\mathbf{z} = (z_1, \dots, z_m)$ of $\mathbb{R}^m$, $m \geq 2$, we first define the respective weak, strict and strong dominance operators $\preccurlyeq$, $\precsim$ and $\prec$ as:

$$\begin{cases} \mathbf{z} \preccurlyeq \mathbf{y} & \Leftrightarrow \quad \forall i = 1, \dots, m,\ z_i \leq y_i\,, \\ \mathbf{z} \precsim \mathbf{y} & \Leftrightarrow \quad \forall i = 1, \dots, m,\ z_i \leq y_i \ \text{ and } \ \exists i \in \{1, \dots, k\},\ z_i < y_i\,, \\ \mathbf{z} \prec \mathbf{y} & \Leftrightarrow \quad \forall i = 1, \dots, m,\ z_i < y_i\,. \end{cases}$$

The expression *weak dominance* is used here as in [54, Section 14.2], or [36], *strict* dominance as in [8], Definition 2.1, and *strong dominance* as in [9, Section 2.4.5]. *Strict dominance* is usually referred simply as *dominance* or *Pareto dominance*. Notice that the terminology or symbols employed differ among authors.