Incremental Gaussian Model-Building in Multi-Objective EDAs with an Application to Deformable Image Registration

Peter A.N. Bosman
Centrum Wiskunde & Informatica (CWI)
P.O. Box 94079
1090 GB Amsterdam
The Netherlands
Peter.Bosman@cwi.nl

Tanja Alderliesten
The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital (NKI-AVL)
P.O. Box 90203
1006 BE Amsterdam
The Netherlands
T.Alderliesten@nki.nl

ABSTRACT

Estimation-of-Distribution Algorithms (EDAs) build and use probabilistic models during optimization in order to automatically discover and use an optimization problems' structure. This is especially useful for black-box optimization where no assumptions are made on the problem being solved, which is characteristic of many cases in solving complex real-world problems. In this paper we consider multi-objective optimization problems with real-valued variables. Although the vast majority of advances in EDA literature concern single-objective optimization, advances have also been made in multi-objective optimization. In this paper we bring together two recent advances, namely incremental Gaussian model building to reduce the required population size and a mixture-based multi-objective framework that has specific methods to better facilitate model-building techniques that span multiple generations. Significantly faster convergence to the optimal Pareto front is achieved on 6 out of 7 artificial benchmark problems from literature. Although results on two of these problems show that building models with higher-order interactions between variables is required, these problems are still artificial. We therefore also consider a more realistic optimization problem in image processing, namely deformable image registration. For this problem too, our results show the need for processing interactions between problem variables, stressing the importance of studying the use of such models. Furthermore, the number of problem variables in the deformable image registration problem can be very large. The building of models with higher-order interactions, especially mixture-based models, then requires very large population sizes. The use of incremental model building is therefore of high importance. This claim is supported by our results that show a huge reduction in the number of required evaluations on this problem.

GECCO'12, July 7–11, 2012, Philadelphia, Pennsylvania, USA.
Copyright 2012 ACM 978-1-4503-1177-9/12/07 ...$10.00.

Categories and Subject Descriptors

G.1 [**Numerical Analysis**]: Optimization; I.2 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms

Algorithms, Performance, Experimentation

Keywords

Evolutionary Algorithms, Estimation-of-Distribution Algorithms, Multi-Objective Optimization, Incremental Model Building, Deformable Image Registration

1. INTRODUCTION

Many optimization problems in practice are actually multi-objective optimization (MO) problems [6]. In such problems more than one objective functions must be optimized simultaneously without an expression of weights or other means of scalarizing these objectives. A typical example is the design of a product where one objective is the quality of the design and another objective is the associated cost. Minimizing the cost will in general lead to products of inferior quality, while maximizing the quality will give rise to an increasing cost. As a consequence, there are many optimal solutions to a MO problem, all of which represent a trade-off between the two objectives. The notion of searching a search space through maintaining a set of solutions is a key characteristic of evolutionary algorithms (EAs). EAs are commonly accepted to be well-suited for solving MO problems [6]. Because a set of solutions is used, EAs can spread their search bias along the Pareto front and thereby prevent many re-computations that are involved if a single point on the Pareto front is repeatedly targeted using a single-solution based approach.

The efficiency of variation operators when used in a multi-objective EA (MOEA) can often be improved by employing restricted mating such that solutions that are closer to each other in objective space have a higher probability of being combined. In EDAs a natural way to achieve this is by building mixture distributions that spread the mixture components throughout objective space [4, 8, 11]. Such an approach allows EDAs to spread the search intensity along the Pareto front, allowing more focused exploitation of problem structure in different regions of the objective space.

Recently, a specific way of building and using mixture models in EDAs was introduced that can be used to extend single-objective EDAs to MO [4]. Particular to this work is that it explicitly allows mixture components to overlap in the objective space in order to cover the entire Pareto front and prevent convergence of mixture components to singular points on the front. Furthermore, the progression of mixture components in objective space is explicitly tracked so as to increase the efficiency of adaptive distribution enhancement techniques that span over multiple generations such as adaptive variance scaling techniques [4] to prevent premature convergence. The downside to this approach to using mixture distributions is that for a mixture distribution with k mixture components, a population size is required that is $\mathcal{O}(k)$ times larger. Especially when building higher-order probabilistic models, as is common in EDA literature, the required population size tends to be quite large. The additional requirement imposed by the aforementioned approach to using mixture distributions for MO then can become problematic. However, some studies have shown that probabilistic models can be learned effectively in EDAs over multiple generations [3, 9]. This can greatly reduce the required population size. Such an approach therefore is a good candidate to reduce the number of required evaluations in mixture-based EDAs for MO. Moreover, the recent work on tracking mixture components across generations fits perfectly with the idea of building models incrementally over multiple generations. Therefore, in this paper we combine incremental model building with this approach, and experimentally validate the performance resulting from the reduction in required population size. Specifically, we consider the use of a recent Gaussian EDA named AMaLGaM [3] and its incremental model-building counterpart, iAMaLGaM.

In addition to experimentally validating the merits of EDAs on well-understood and well-known artificial benchmark problems, it is important to also consider non-artificial benchmark problems. Besides obtaining realistic insights about the potential real-world performance of EDAs, it is important to validate issues such as whether key research efforts in EDA literature, such as the building of probabilistic models that cover interactions between problem variables, is really needed and leads to better performance in practice as well. In addition to well-known artificial benchmark problems, in this paper we therefore also consider a non-artificial problem, namely deformable image registration.

The remainder of this paper is organized as follows. In Section 2 we provide an overview of the most important concepts and definitions in multi-objective optimization. Then, in Section 3 we outline the single-objective AMaLGaM and the multi-objective mixture-based EDA framework. In Section 4 we discuss the iAMaLGaM approach to incremental Gaussian model building and its use in the multi-objective mixture-based EDA framework. We perform an experimental analysis on 7 well-known benchmark problems from the literature and a deformable image registration problem in Section 5 and draw our final conclusions in Section 6.

2. MULTI-OBJECTIVE OPTIMIZATION

We assume to have m objective functions $f_i(\boldsymbol{x})$, $i \in \{0, 1, \dots, m-1\}$ and, without loss of generality, we assume that the goal is to *minimize* all objectives.

A solution $\boldsymbol{x}^0$ is said to (Pareto) *dominate* a solution $\boldsymbol{x}^1$ (denoted $\boldsymbol{x}^0 \succ \boldsymbol{x}^1$) if and only if $f_i(\boldsymbol{x}^0) \leq f_i(\boldsymbol{x}^1)$ holds for all $i \in \{0, 1, \dots, m-1\}$ and $f_i(\boldsymbol{x}^0) < f_i(\boldsymbol{x}^1)$ holds for at least one $i \in \{0, 1, \dots, m-1\}$. A *Pareto set* of size n then is a set of solutions $\boldsymbol{x}^j$, $j \in \{0, 1, \dots, n-1\}$ for which no solution dominates any other solution, i.e. there are no $j, k \in \{0, 1, \dots, n-1\}$ such that $\boldsymbol{x}^j \succ \boldsymbol{x}^k$ holds. A *Pareto front* corresponding to a Pareto set is the set of all m-dimensional objective function values corresponding to the solutions, i.e. the set of all $\boldsymbol{f}(\boldsymbol{x}^j)$, $j \in \{0, 1, \dots, n-1\}$.

A solution $\boldsymbol{x}^0$ is said to be *Pareto optimal* if and only if there is no other $\boldsymbol{x}^1$ such that $\boldsymbol{x}^1 \succ \boldsymbol{x}^0$ holds. Further, the *optimal Pareto set* is the set of all Pareto-optimal solutions and the *optimal Pareto front* is the Pareto front that corresponds to the optimal Pareto set. We denote the optimal Pareto set by $\boldsymbol{\mathcal{P}_S}$ and the optimal Pareto front by $\boldsymbol{\mathcal{P}_F}$.

3. GENERATIONAL MODEL-BUILDING

In this section we briefly describe the single-objective EDA that we consider as well as the extension thereof to MO via a specific use of mixture distributions.

3.1 AMaLGaM

The Adapted Maximum-Likelihood Gaussian Model (AMaLGaM) is a Gaussian EDA that estimates a Gaussian distribution from the selected solutions with maximum likelihood (ML) and subsequently adaptively changes this estimate on the basis of observations regarding improvements that were found after sampling the Gaussian. For details about these adaptive techniques, we refer the interested reader to the literature [3]. Here we only point out that the main parameters to be estimated in a Gaussian EDA are the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, or, if the distribution is factorized so that the variables are all considered to be independent, the variance vector. Moreover, ML parameter estimates for the Gaussian distribution are well-known. Let $\boldsymbol{\mathcal{S}}$ denote a vector of data. A ML estimation for parameters of the Gaussian probability distribution is obtained from $\boldsymbol{\mathcal{S}}$ if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated by the sample average and sample covariance matrix respectively [1]:

$$\hat{\boldsymbol{\mu}} = \frac{1}{|\boldsymbol{\mathcal{S}}|}\sum_{j=0}^{|\boldsymbol{\mathcal{S}}|-1}(\boldsymbol{\mathcal{S}}_j), \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{|\boldsymbol{\mathcal{S}}|}\sum_{j=0}^{|\boldsymbol{\mathcal{S}}|-1}((\boldsymbol{\mathcal{S}}_j) - \hat{\boldsymbol{\mu}})((\boldsymbol{\mathcal{S}}_j) - \hat{\boldsymbol{\mu}})^T.$$

Although the variance is adaptively scaled up in AMaLGaM to prevent premature convergence, the use of ML estimates aligns the principle axis of variance with the density contours of the search space because of selection. It is however in the perpendicular direction that the most improvement can be obtained in the local fitness landscape. To overcome this misalignment problem, the Anticipated Mean Shift is used in AMaLGaM [3]. The AMS is computed as the difference between the means of subsequent generations, i.e. $\hat{\boldsymbol{\mu}}^{\text{Shift}}(t) = \hat{\boldsymbol{\mu}}(t) - \hat{\boldsymbol{\mu}}(t-1)$. A part of the newly sampled solutions is then moved in the direction of the AMS: $\boldsymbol{x} \leftarrow \boldsymbol{x} + 2\hat{\boldsymbol{\mu}}^{\text{Shift}}(t)$.

3.2 MAMaLGaM-X

A specific extension of single-objective EDAs to MO on the basis of mixture distributions was recently studied and applied to AMaLGaM, resulting in the Multi-objective AMaLGaM-miXture (MAMaLGaM-X) [4]. Using a clustering technique, the selected solutions are clustered into k overlapping clusters of identical size. Specifically, for an overall population size n and selection size τn where $\tau \in [\frac{1}{n}; 1]$, each cluster will have $(2\tau n)/k$ solutions. Subsequently, distribution estimation and sampling proceeds as normally done in the EDA, but independently per cluster.

The problem with solutions aligning with contours of the fitness landscape also exists in MO [4]. For this reason, techniques such as AMS are important in MO as well. However, to ensure that AMS works well, there needs to be a meaningful progression in the Gaussian distribution in subsequent generations. In the mixture distribution this means that individual clusters need to be associated with each other in a sensible manner across subsequent generations. This is however not automatically the case by clustering the selected solutions anew in each generation. Therefore, in MAMaLGaM-X an explicit cluster registration step is performed that computes the overall most likely association between clusters of subsequent generations by minimizing the sum of all distances between associated clusters [4].

An elitist archive is maintained, storing all currently non-dominated solutions. Because the objectives are real-valued, there are typically infinitely many non-dominated solutions possible. To prevent the archive from growing to an extreme size, the objective space is discretized into hypercubes by discretizing each objective separately. Only one solution per hypercube is allowed in the archive. Newly generated solutions are compared to the solutions in the archive. If a new solution is dominated by any archive solution, it is not entered. If a new solution is not dominated, it is added to the archive if the hypercube that it resides in does not already contain a solution or if it dominates that particular solution. When a new solution is entered, all archive solutions that are dominated by it, are removed.

3.3 MAMaLGaM-X$^+$

An extension of MAMaLGaM-X called MAMaLGaM-X$^+$ was shown to exhibit the most robust performance on a set of well-known benchmark problems. The difference between MAMaLGaM-X$^+$ and MAMaLGaM-X is that m additional clusters are maintained, one for each objective. Selection in these clusters is done completely independently on the basis of each respective individual objective, thereby specifically targeting convergence at the extreme regions of the Pareto front. Solutions from these specific clusters are furthermore also integrated into the selection procedure on the basis of which the k clusters are computed for the mixture distribution in MAMaLGaM-X. For details, see [4].

4. INCREMENTAL MODEL-BUILDING

Estimating the probability distribution in AMaLGaM and in MAMaLGaM-X is done anew from scratch each generation. However, subsequent generations have much in common, which allows the required population size to be reduced using incremental learning, i.e. combining the probability distribution that is estimated from the selected solutions in this generation with the distribution that was used in the previous generation. A specific approach to doing this is provided in the incremental AMaLGaM (iAMaLGaM).

4.1 iAMaLGaM

In iAMaLGaM a memory-decay approach is taken for the covariance matrix and the AMS. The equations for incrementally estimating the covariance matrix and the AMS in generation t are:

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}(t) \quad &= \quad (1-\eta^{\boldsymbol{\Sigma}})\hat{\boldsymbol{\Sigma}}(t-1)+ \\ &\quad\;\; \eta^{\boldsymbol{\Sigma}} \tfrac{1}{|\boldsymbol{\mathcal{S}}|}\textstyle\sum_{i=0}^{|\boldsymbol{\mathcal{S}}|-1} \left(\boldsymbol{\mathcal{S}}_i-\hat{\boldsymbol{\mu}}(t)\right)\left(\boldsymbol{\mathcal{S}}_i-\hat{\boldsymbol{\mu}}(t)\right)^T \\ \hat{\boldsymbol{\mu}}^{\text{Shift}}(t) \quad &= \quad (1-\eta^{\text{Shift}})\hat{\boldsymbol{\mu}}^{\text{Shift}}(t-1)+ \\ &\quad\;\; \eta^{\text{Shift}}\left(\hat{\boldsymbol{\mu}}(t)-\hat{\boldsymbol{\mu}}(t-1)\right).\end{aligned}$$

Values for the learning-rate parameters $\eta^{\boldsymbol{\Sigma}}$ and η^{Shift} were determined empirically [3].

4.2 iMAMaLGaM-X$^{(+)}$

We combine the use of incremental model learning with the specific mixture-based EDA extension as outlined above. This means that for every cluster, including the ones in which selection is done separately on the basis of individual objectives (e.g. as in MAMaLGaM-X$^+$), we have a separate incremental Gaussian model that is updated using the incremental updates provided above. The explicit cluster registration in MAMaLGaM-X was shown to work well in combination with the AMS technique. It is therefore expected that the so-established relation between clusters in subsequent generations is sufficiently meaningful for incremental model building to work as well. We will identify the resulting multi-objective EDAs as incremental MAMaLGaM-X (iMAMaLGaM-X) and incremental MAMaLGaM-X$^+$ (iMAMaLGaM-X$^+$).

5. EXPERIMENTS

In this section we present our experimental analysis of iMAMaLGaM-X and iMAMaLGaM-X$^+$. First, we describe the optimization problems that we consider, both the common benchmark problems and the deformable image registration problem. Next, we describe how we evaluate performance and then we present our results.

5.1 Common Benchmark Problems

The definitions of the problems in our multi-objective optimization problem test suite are presented in Table 1.

We used the well-known problems[1] EC$_i$, $i \in \{1, 2, 3, 4, 6\}$. The initialization ranges (IRs) of the EC$_i$ problems are also hard constraints. The reason for this is that otherwise some objectives can not always be evaluated. Such rigid constraints can be hard for a numerical optimizer. EC$_1$ and EC$_2$ are continuous and do not have any local fronts. EC$_1$ has a convex Pareto front whereas EC$_2$ has a concave Pareto front. These problems differ from the GM problems in that the objectives are not similarly defined and not similarly scaled. EC$_3$ is similar to EC$_1$ but has a discontinuous Pareto front. EC$_4$ has many locally optimal Pareto fronts. Finally, the Pareto front of EC$_6$ is non-uniformly distributed. For more details about these functions, see [12].

Two additional problems come from more recent literature on real-valued MO optimization [5] and are labeled BD$_i$, $i \in \{1, 2\}$. Both problems make use of Rosenbrock's function. Premature convergence on this function is likely without proper induction of the structure of the search space. Function BD$_2$ is harder than BD$_1$ in that the objective functions overlap in all variables instead of only in x_0. Further, the IR of x_0 in function BD$_1$ is also a constraint. Finally, we

[1] These problems are also known as ZDT$_i$.

have scaled the objectives of BD$_2$ to ensure that the optimum of all problems is in approximately the same range. By doing so, using the same value-to-reach for the $\boldsymbol{D_{\mathcal{P}_F \rightarrow \mathcal{S}}}$ indicator (which is explained in the next Section) on all problems corresponds to a similar front quality on all problems.

Name	Objectives	IR
EC$_1$	$f_0 = x_0, \quad f_1 = \gamma\left(1 - \sqrt{f_0/\gamma}\right)$ $\gamma = 1 + 9\left(\sum_{i=1}^{l-1} x_i/(l-1)\right)$	$[0;1]^{30}$ $(l = 30)$
EC$_2$	$f_0 = x_0, \quad f_1 = \gamma\left(1 - (f_0/\gamma)^2\right)$ $\gamma = 1 + 9\left(\sum_{i=1}^{l-1} x_i/(l-1)\right)$	$[0;1]^{30}$ $(l = 30)$
EC$_3$	$f_0 = x_0, \quad f_1 = \gamma\left(1 - \sqrt{f_0/\gamma} - (f_0/\gamma)\sin(10\pi f_0)\right)$ $\gamma = 1 + 9\left(\sum_{i=1}^{l-1} x_i/(l-1)\right)$	$[0;1]^{30}$ $(l = 30)$
EC$_4$	$f_0 = x_0, \quad f_1 = \gamma\left(1 - \sqrt{f_0/\gamma}\right)$ $\gamma = 1 + 10(l-1) + \sum_{i=1}^{l-1}\left(x_i^2 - 10\cos(4\pi x_i)\right)$	$[-1;1]\times$ $[-5;5]^9$ $(l = 10)$
EC$_6$	$f_0 = 1 - e^{-4x_0}\sin^6(6\pi x_0), \quad f_1 = \gamma\left(1 - (f_0/\gamma)^2\right)$ $\gamma = 1 + 9\left(\sum_{i=1}^{l-1} x_i/(l-1)\right)^{0.25}$	$[0;1]^{10}$ $(l = 10)$
BD$_1$	$f_0 = x_0, \quad f_1 = 1 - x_0 + \gamma$ $\gamma = \sum_{i=1}^{l-2}\left(100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2)\right)$	$[0;1]\times$ $[-5.12;5.12]^9$ $(l = 10)$
BD$_2^{\mathrm{s}}$	$f_0 = \frac{1}{l}\sum_{i=0}^{l-1} x_i^2$ $f_1 = \frac{1}{l-1}\sum_{i=0}^{l-2}\left(100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2)\right)$	$[-5.12;5.12]^{10}$ $(l = 10)$

Table 1: The MO problem test suite.

It is important to note that for the EC$_i$ problems and for x_0 in function BD$_1$, this IR is a constraint. If these variables move outside of their IR, some objective values can become non-existent. It is therefore important to keep these variables within their IRs. However, a simple repair mechanism that changes a variable to its boundary value if it has exceeded this boundary value gives artifacts. If for instance the search on problem EC$_6$ probes a solution that has a negative value for each of the variables x_i with $i \geq 1$, then the repair mechanism sets all these variables to 0. The solution that results after boundary repair lies on the Pareto front. To avoid artifacts resulting from boundary-repair methods, the sampling procedure in all MOEDAs is constructed such that solutions that are out of bounds are rejected.

5.2 Deformable Image Registration

Medical imaging is of great value in healthcare. There are a variety of imaging modalities available nowadays (e.g. CT, MRI, PET, etc.), all with different purposes and characteristics. When multiple images are available, either acquired with different imaging techniques or using the same technique but acquired at different points in time, for example for follow-up purposes, important information about the state of a patient lies in comparing and relating these images to one another in order to obtain a more complete picture or see how things changed over time.

The general idea of image registration is to find a transformation that transforms a source image to a target image. By means of rigid registration, only rigid transformations such as rotations and translations are considered. However, for medical applications, *deformable* image registration is of far greater value. This is because in many cases the anatomy that is imaged has changed for various reasons, including changes as a result of treatment, disease progression or simply a difference in patient orientation during image aquisition, resulting in anatomical changes due to different effects of gravity. Although deformable image registration is of great value, it is also generally a hard problem.

For the task of deformable image registration two issues are of prime interest: 1) intensity similarity, i.e. the degree of similarity between intensity patterns in the target image and the transformed source image, and 2) transformation effort, i.e. the amount of energy required to accomplish the transformation. Even within specific real-world problems such as this, competent BBO optimizers such as (most) EDAs can play an important role. Not only is the problem multi-objective, which already makes a set-based approach such as EDAs of interest, also there are many different ways possible to compute similarities and there are many transformation models possible. For more advanced transformation models, as typically required for deformable image registration, optimization already becomes harder to do problem-specifically, but moreover, it is undesirable to have to laboriously design specific optimization algorithms for each different combination of transformation model and similarity model. It is however not the main purpose of this paper to provide a rigorous study into different models for image registration. Therefore, in the following we provide rudimentary, but computationally useful models.

5.2.1 Representation

The transformation model, i.e. the representation of possible transformations, is a square grid of size $n_g \times n_g$. The grid overlays the source image in a regular manner, meaning that it corresponds to a subdivision of the source image into $(n_g - 1)(n_g - 1)$ equally-sized axes-parallel rectangles (see Figure 1). The actual transformation then is given by the association of coordinates with each point in the grid. A means of interpolation is required to extend the so-established correspondence between grids to create the transformed source image. We use bi-linear interpolation in each rectangle.

5.2.2 Similarity measure

We model similarity in intensity with a measure (to be maximized) that is commonly adopted in registration literature, namely normalized mutual information:

$$\frac{(H(T[s]) + H(t))}{H(T[s], t)} - 1,$$

where $H(T[s])$, $H(t)$ and $H(T[s], t)$ denote the entropy of the probability distribution of the grey values in the transformed source image, the entropy of the probability distribution of the grey values in the target image and the entropy of the joint probability distribution of the grey values (i.e. for the registered pairs of pixels) in the transformed source image and the target image, respectively [10]. Because we will consider minimization of all objectives, we will consequently minimize the negative normalized mutual information.

5.2.3 Deformation energy

To associate physical characteristics with transformations, Hooke's law is used [2]. Because we are interested in non-rigid transformations, transformations such as rotations and translations of the entire grid should not correspond to an increase in energy. The required energy is therefore computed on the basis of changes in the lengths of edges in the grid. To ensure that the physical changes we are interested in, i.e. non-rigid deformations of subrectangles, are always

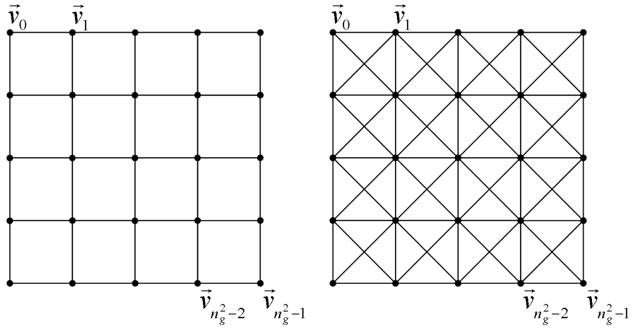

Figure 1: Left: grid of points used as a basis for the transformation model. Right: grid of points with all connections taken into account in the calculation of the required energy to accomplish a transformation.

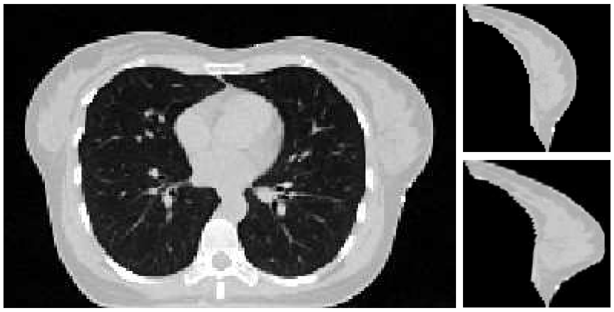

Figure 2: Left: axial slice of a CT scan. Right top: segmented left breast (source image). Right bottom: artificially deformed left breast (target image).

associated with an increase in required energy, we also include the diagonal edges in each subrectangle (see Figure 1). Now, if we denote the grid coordinates in the source and target images by vectors $\boldsymbol{v}_i^{before}, \boldsymbol{v}_i^{after}, i \in \{0, 1, \ldots, n_g^2 - 1\}$ and the set of considered edges by E, we can define the total energy $U_{total\text{-}deform}$ to be minimized as follows:

$$U_{total\text{-}deform} = \sum_{(i,j)\in E} U_{deform}(i,j)$$

where

$$U_{deform}(i,j) =$$

$$\frac{1}{2} l_{ij} \left(\| \boldsymbol{v}_i^{before} - \boldsymbol{v}_j^{before} \| - \| \boldsymbol{v}_i^{after} - \boldsymbol{v}_j^{after} \| \right)^2$$

where l_{ij} is an elasticity constant associated with the tissue that edge (i, j) crosses.

5.2.4 *Optimization*

In theory, the goal now becomes to find the transformation that corresponds to minimal energy while obtaining perfect similarity between the images, i.e. a constrained single-objective optimization problem. However, in practice, due to issues such as noise in image acquisition, inaccuracy in the determination of the parameters of the physical model (i.e. segmentation and values for material properties) and the inability to mathematically compactly represent all possible transformations, a transformation that results in perfect similarity may not exist. Moreover, transformations that result in a larger similarity are not necessarily preferable. Therefore, the underlying problem in practice is actually multi-objective, i.e. find transformations that on the one hand maximize the similarity between source and target image (objective 1) and on the other hand minimize the amount of required energy (objective 2). In addition to these objectives, constraints are required that are well-known to be important in deformable image registration such as prohibiting transformations that fold the grid. We incorporated these constraints into the multi-objective optimization approach using constraint domination by means of which a solution is always preferred in selection if it doesn't violate any constraints [7]. Finally, we used a single problem instance in terms of pairs of images to perform deformable image registration on with $n_g = 5$, giving a 50-dimensional real-valued constrained multi-objective optimization problem. The pair of images to be registered is shown in Figure 2. Both images have a dimensionality of 125×125 pixels.

5.3 Measuring performance

We consider the Pareto set that can be computed from the elitist archive combined with the population upon termination to be the outcome of running an EDA and refer to it as an approximation set, denoted $\mathcal{S}$. To measure performance the $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F}\to\mathcal{S}}$ performance indicator is computed. This performance indicator computes the average distance over all points in the optimal Pareto front $\boldsymbol{\mathcal{P}_F}$ to the nearest point in $\mathcal{S}$:

$$\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F}\to\mathcal{S}}(\mathcal{S}) = \frac{1}{|\boldsymbol{\mathcal{P}_F}|} \sum_{\boldsymbol{f}^1\in\boldsymbol{\mathcal{P}_F}} \min_{\boldsymbol{f}^0\in\mathcal{S}} \{d(\boldsymbol{f}^0, \boldsymbol{f}^1)\}$$

where $\boldsymbol{f}$ is a point in objective space and $d(\cdot,\cdot)$ computes Euclidean distance. A smaller $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F}\to\mathcal{S}}$ value is preferable and a value of 0 is obtained if and only if the approximation set and the optimal Pareto front are identical. This indicator is useful for evaluating performance if the optimum is known because it describes how well the optimal Pareto front is covered and thereby represents an intuitive trade-off between the diversity of $\mathcal{S}$ and its proximity (i.e. closeness to the optimal Pareto front). Even if all points in $\mathcal{S}$ are on the optimal Pareto front the indicator is not minimized unless the solutions in the approximation set are spread out perfectly. Because the optimal Pareto front may be continuous, there are infinitely many solutions possible on the optimal Pareto front. Therefore, we computed 5000 uniformly sampled solutions along the optimal Pareto front to use as a discretized version of $\boldsymbol{\mathcal{P}_F}$ for a high-quality approximation.

For the problems in our test-suite, given the ranges of the objectives for the optimal Pareto front configurations, a value of 0.01 for the $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F}\to\mathcal{S}}$ indicator corresponds to fronts that are quite close to the optimal Pareto front. Fronts that have a $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F}\to\mathcal{S}}$ value of 0.01 can be seen in Figure 3.

5.4 Results

MAMaLGaM-X, MAMaLGaM-X$^+$, iMAMaLGaM-X and iMAMaLGaM-X$^+$ were run using an unfactorized Gaussian, i.e. with a full covariance matrix, and a univariately factorized Gaussian, i.e. with all variables modeled as independent. All presented results are averaged over 10 independent runs. The subpopulation or cluster sizes were set according to guidelines from recent literature on SO [3]:

- Full covariance matrix
  generational: $n^{sub} = 17 + 3l^{1.5}$
  incremental: $n^{sub} = 10l^{0.5}$
- Univariate factorization
  generational: $n^{sub} = 10l^{0.5}$
  incremental: $n^{sub} = 4l^{0.5}$

The overall population size is $\frac{1}{2}k$ times the cluster size. Furthermore, we used $k = 20$ clusters as this number was previously found to provide excellent results [4]. The discretization of the objectives into hypercubes for the elitist archive is set to 10^{-2}.

We observe the average convergence of the $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F} \rightarrow \mathcal{S}}$ metric to study the impact of incremental model building in combination with estimating Gaussian distributions. The average convergence results are shown in Figure 4. It is clear from the results that overall, iMAMaLGaM-X$^+$ performs the best as it converges to the lowest $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F} \rightarrow \mathcal{S}}$ score on all problems except EC$_4$ and does so with the least number of function evaluations, with the exception of problem BD$_2^s$. Moreover, the higher the dimensionality of the problem, the bigger the difference. This is the clearest on the deformable registration (DR) problem, which has the largest dimensionality of all problems in our test suite, when the full covariance matrix is used. The difference in required number of function evaluations before convergence is huge. In the case of a problem like DR this is of prime importance because function evaluations take much more time than is the case for the artificial benchmark problems. Consequently, using MAMaLGaM-X$^{(+)}$ on DR with the full covariance matrix takes hours to run whereas with the incremental variants useful answers are already found within 30 minutes.

The results also show that the use of incremental model building narrows the gap between use of the full covariance matrix and the univariate factorization in the mixture-based MO setting. However, on problems where modelling dependencies between problem variables is not required (i.e. on EC$_1$, EC$_2$, EC$_3$ and EC$_6$, use of the univariately factorized Gaussian distribution is still far more efficient. This indicates that there is still a performance gain to be achieved by online deciding whether to process dependencies.

Finally, the results show that modeling dependencies can be of importance to reliably converge to the optimal Pareto front. On both problem BD$_1$ and BD$_2^s$ the use of a univariately factorized Gaussian distribution leads to convergence to worse $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F} \rightarrow \mathcal{S}}$ values, although this convergence does proceed much faster for incremental model building on the BD$_2^s$ problem. A similar observation can be made for the DR problem. Although it is not easy to see in the logarithmic convergence graphs in Figure 4, the univariate models all converge to worse $\boldsymbol{D}_{\boldsymbol{\mathcal{P}_F} \rightarrow \mathcal{S}}$ values than do their unfactorized counterparts. This can be better observed in Figure 5 where the Pareto front is shown computed for each algorithm over all runs. The difference between the unfactorized Gaussian models and the factorized Gaussian models is clear as there is a part of the Pareto front, corresponding to larger deformations of the grid and better similarity values, that cannot be reached (efficiently) by the univariate models. This stresses the need for considering dependencies also in real-world problems.

For completeness, in Figure 6 we show the transformation that was found with the best similarity value. Given the degrees of freedom in the transformation model, a perfect match cannot be obtained, but the match is already close enough to be of real-world clinical relevance. Considering that a black-box approach to solving this problem was taken, this further illustrates how EDAs can find high-quality solutions and make a solid contribution to real-world problem solving. Also, in image registration literature, a linear combination is always made of the two objectives. Not only is there no theoretical underpinning available for selecting a weight for such as a linear combination, it is known from multi-objective optimization theory that if part of the Pareto front is concave, no solutions on this part of the Pareto front can be found using a weighted linear combination [6], which speaks in favor of using a multi-objective approach instead, like the one studied in this paper.

6. CONCLUSIONS

In this paper, we have combined the use of Gaussian mixture distributions in an EDA for multi-objective optimization with an approach to incremental model building spanning multiple generations. The use of mixture distributions is very useful in multi-objective optimization as it allows spreading the search bias of the optimization algorithm across the Pareto front in a natural manner. We found that using mixture-component registration that associates the nearest mixture components in subsequent generations with each other, incremental model-building techniques that span multiple generations can successfully be applied. The reduction in required population size is k-fold larger compared to the single-objective case if k mixture components are used. As a result, we observed a substantial reduction in the required number of function evaluations on a set of common multi-objective benchmark problems as well as on a higher-dimensional multi-objective application problem: deformable image registration. Taken over all results, the multi-objective EDA that was introduced as iMAMaLGaM-X$^+$ (incremental Multi-objective AMaLGaM-miXture with parallel single-objective optimization clusters) was found to be the most promising variant to consider in future research.

7. REFERENCES

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons Inc., New York, 1958.

[2] G. Arfken. *Mathematical methods for physicists*. Academic Press, Inc., San Diego, 1985.

[3] P. A. N. Bosman. On empirical memory design, faster selection of Bayesian factorizations and parameter-free Gaussian EDAs. In G. Raidl et al., editors, *Proc. of the Genetic and Evolutionary Comp. Conf. — GECCO–2009*, pages 389–396, New York, 2009. ACM Press.

[4] P. A. N. Bosman. The anticipated mean shift and cluster registration in mixture-based EDAs for multi-objective optimization. In J. Branke et al., editors, *Proc. of the Genetic and Evolutionary Comp. Conf. — GECCO–2010*, pages 351–358, New York, 2010. ACM Press.

[5] P. A. N. Bosman and D. Thierens. Adaptive variance scaling in continuous multi–objective estimation–of–distribution algorithms. In D. Thierens et al., editors, *Proceedings of the Genetic and Evolutionary Computation Conference — GECCO–2007*, pages 500–507, New York, 2007. ACM Press.

[6] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons Inc., New York, 2001.

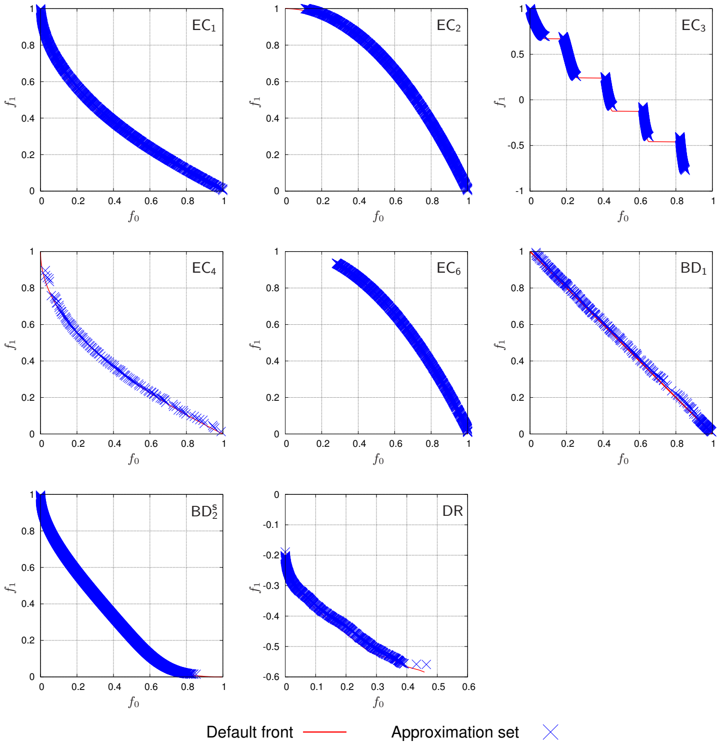

Figure 3: Default fronts and approximation sets obtained with iMAMaLGaM-X$^+$ ($D_{\mathcal{P}_F \to \mathcal{S}} = 0.01, k = 20$).

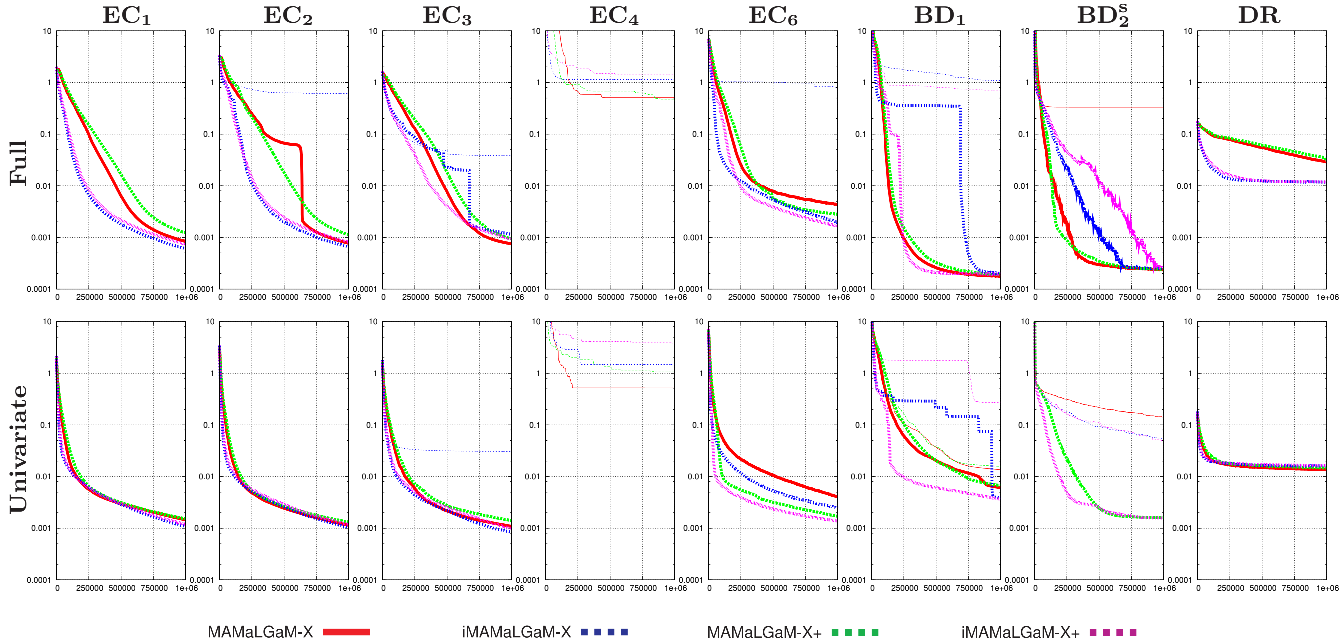

Figure 4: Average performance of various MOEDAs on all problems, estimating full covariance matrices in each cluster. Horizontal axis: number of evaluations (both objectives per evaluation). Vertical axis: $D_{\mathcal{P}_F \to \mathcal{S}}$. For each algorithm averages are shown both for successful runs (bold) and unsuccessful runs, giving double occurrences of lines if some runs were unsuccessful.

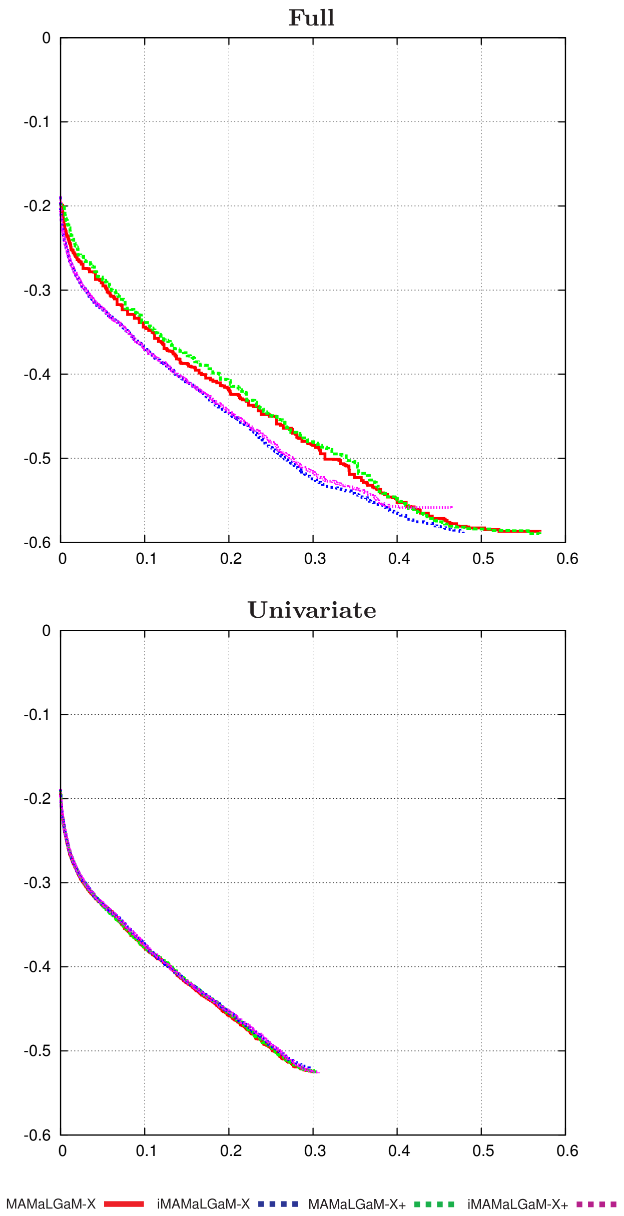

Figure 5: Pareto fronts computed over all runs on the deformable image registration problem instance. Horizontal axis: energy objective. Vertical axis: similarity objective.

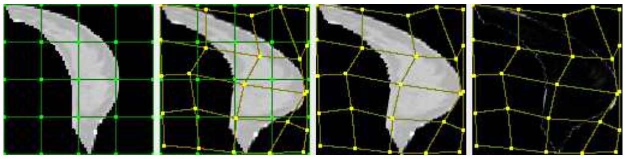

Figure 6: Best similarity result for the deformable image registration problem instance. From left to right: source image, target image, transformed source image, absolute difference of target and transformed source image. Green: initial regular grid. Yellow: transformed grid.

[7] K. Deb, A. Pratap, and T. Meyarivan. Constrained test problems for multi–objective evolutionary optimization. In E. Zitzler et al., editors, *Evolutionary Multi–Criterion Optimization — EMO–2001*, pages 284–298, Berlin, 2001. Springer–Verlag.

[8] M. Pelikan, K. Sastry, and D. E. Goldberg. Multiobjective hBOA, clustering, and scalability. In H.-G. Beyer et al., editors, *Proceedings of the Genetic and Evolutionary Computation Conference — GECCO–2005*, pages 663–670, New York, 2005. ACM Press.

[9] M. Pelikan, K. Sastry, and D. E. Goldberg. iBOA: the incremental bayesian optimization algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference — GECCO–2008*, pages 455–462, New York, 2008. ACM Press.

[10] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans Med Imag*, 22:986–1004, 2004.

[11] D. Thierens and P. A. N. Bosman. Multi–objective mixture–based iterated density estimation evolutionary algorithms. In L. Spector et al., editors, *Proceedings of the Genetic and Evolutionary Computation Conference — GECCO–2001*, pages 663–670, San Francisco, California, 2001. Morgan Kaufmann.

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