On the Importance of Diversity Maintenance in Estimation of Distribution Algorithms

Bo Yuan
School of Information Technology and Electrical Engineering
The University of Queensland
QLD 4072, Australia
+61-7-33651636
boyuan@itee.uq.edu.au

Marcus Gallagher
School of Information Technology and Electrical Engineering
The University of Queensland
QLD 4072, Australia
+61-7-33656197
marcusg@itee.uq.edu.au

ABSTRACT
The development of Estimation of Distribution Algorithms (EDAs) has largely been driven by using more and more complex statistical models to approximate the structure of search space. However, there are still problems that are difficult for EDAs even with models capable of capturing high order dependences. In this paper, we show that diversity maintenance plays an important role in the performance of EDAs. A continuous EDA based on the Cholesky decomposition is tested on some well-known difficult benchmark problems to demonstrate how different diversity maintenance approaches could be applied to substantially improve its performance.

Categories and Subject Descriptors
G.1.6 [**Numerical Analysis**]: Optimization –*Global Optimization.*

General Terms
Algorithms, Performance, Experimentation

Keywords
EDAs, Continuous Optimization, Clustering

1. INTRODUCTION
Estimation of Distribution Algorithms (EDAs)[8, 11] refer to a class of novel Evolutionary Algorithms (EAs) based on probabilistic modeling instead of classical genetic operators such as crossover or mutation. The fundamental mechanism is to conduct searching by sampling new individuals from a probability distribution, which is estimated based on some selected promising individuals in the current population. The major advantage of EDAs is that they can explicitly learn the dependences among variables of the problem to be solved and use this structural information to efficiently generate new individuals. It has been shown in previous work that EDAs can outperform traditional EAs on a number of difficult benchmark problems.

GECCO'05, June 25–29, 2005, Washington DC, USA.

Despite of the successful applications of EDAs, an important issue remains: what makes a problem difficult for EDAs and how to solve it? Since EDAs conduct searching based on the structural information learned during evolution, in general, there are two factors that may influence their performance. The first one is whether EDAs are capable of learning the structure of a problem. In continuous spaces, which are the focus of this paper, most EDAs assume that selected individuals can be reasonably approximated by a multivariate Gaussian distribution[3, 7]. However, it is clear that in many cases, for example multimodal landscapes, good individuals are likely to be distant from each other and may not be efficiently represented by one Gaussian distribution. As a result, EDAs may either have to search inefficiently due to the large variances of the Gaussian distribution or stochastically drift towards one optimum and get stuck there. The second factor is whether the global structure of a problem actually leads to the global optimum. Even in some unimodal problems, the global structure may be quite different from the local structure around the global optimum, which means that EDAs utilizing this misleading information may converge to a non-optimal solution.

Recently, there have been some attempts to improve the performance of EDAs in the above situations by using various clustering techniques[3, 9]. The basic idea is to, in each generation, assign selected individuals to a number of clusters first and then apply an EDA on each cluster separately. Although the structure of selected individuals may be too complex to be estimated by one Gaussian distribution, the structure of individuals in each cluster may be much easier to estimate. By using this clustering method, the assumption of one Gaussian distribution can be relaxed to a combination of Gaussian distributions. However, this does not necessarily mean that EDAs with clustering can always achieve satisfactory performance. For example, it has been reported that these EDAs still face much difficulty in the Rosenbrock function, which is unimodal and features a narrow valley towards the global optimum. There are some other issues about the current way of using clustering. For example, in order for clustering techniques to be successful, individuals must present some kind of clustering pattern and this pattern should reflect the underlying structure of the problem (i.e., the multimodality of landscape). However, in some situations, selected individuals may present misleading information.

In the first part of our work, we focus on the analysis of the Rosenbrock function and why it is difficult for EDAs. Furthermore, we show how a simple diversity maintenance method can significantly help EDAs find the global optimum. In

the next part, we propose a novel three-step-method of combining clustering and EDAs. Instead of applying clustering in each generation from the very beginning of evolution, an EA with diversity maintenance technique is employed to conduct rough searching to locate the basins of optima. A clustering technique is then applied on the final population to cluster selected individuals. Finally, an EDA is run on each subset of individuals to conduct fine searching until the optimum is found.

The content of this paper is structured as follows. The framework of EDAs is given in the next Section. Section 3 introduces two benchmark problems to be tested. Experimental results are presented in Section 4 and Section 5 concludes our work and points out some directions for further work.

2. THE FRAMEWORK OF EDAs

The general framework of EDAs is given in Table 1, although they may differ from each other in a number of details. In EDAs, the most important part is how to estimate the probability distribution θ^{sel} of selected individuals.

Table 1. The Framework of EDAs.

```
Initialize and evaluate the population P
  While stopping criteria not met
    Select some individuals P^sel from P
    Estimate the density function θ^sel
    Create P′ by sampling from θ^sel
    Evaluate individuals in P′
    Combine P and P′ to create the new P
  End While
```

Since it is usually difficult to directly sample from a multivariate probability distribution, some EDAs utilize conditional factorization in which a multivariate probability distribution is represented by the product of a set of univariate conditional probability distributions:

$$P(x_1,..,x_n)=\prod_{i=1}^{n}P(x_i \mid x_{i+1},...,x_n) \qquad \text{Eq. 1}$$

However, it may still not be convenient to calculate a conditional probability distribution involving M parents when M is large. For example, in binary spaces, the number of individuals needed to conduct reliable estimation will increase exponentially as M goes up. There is a similar issue in continuous spaces for certain probabilistic models such as histograms in which the number of bins and the number of individuals required will also quickly become extremely large as dimensionality increases. As a result, an upper limit k of the number of parents is usually set, which assumes that the problem to be solved only has limited order of dependences. Since conditional factorizations can be represented by acyclic graphs, a searching method is needed to find a new graph that can approximate the original probability distribution as well as possible with the constraint of the maximum number of parents. A commonly used score metric is the Kullback-Leibler cross-entropy measure [6] specifying the distance between two probability distributions (to be minimized):

$$D_{K-L}(p(x),g(x))=\int p(x)\cdot\log\frac{p(x)}{g(x)}dx \qquad \text{Eq. 2}$$

However, if the order of dependences in the problem is beyond k, EDAs may not work well because they cannot capture all necessary dependences. Also, finding a factorization can be a time-consuming process and usually the resulting graph is a sub-optimal solution[2].

In this paper, we will use a simple EDA, which is based on a multivariate Gaussian distribution and does not need to calculate conditional probability distributions. Suppose that selected individuals follow an N-D Gaussian (μ, $\sum$). There are N parameters specifying the mean vector μ and $N\cdot(N+1)/2$ parameters specifying the covariance matrix $\sum$, which can be conveniently estimated from the current selected population using their maximum likelihood estimates. By using a Cholesky decomposition, it is easy to find an N-by-N lower triangular matrix S subject to the condition of $\sum=SS^T$[12]. New individuals can be sampled from this Gaussian by:

$$X=\mu+S\cdot Z \qquad \text{Eq. 3}$$

where Z is an N-by-P matrix with each element independently drawn from a Gaussian distribution G (0,1) and P is the number of new individuals[14]. It has been shown that this class of EDAs performed very well compared to more sophisticated EDAs [13].

3. TEST PROBLEMS

3.1 The Rosenbrock Function

The n-dimensional Rosenbrock function is given by:

$$f_{Rosen}(X)=\sum_{i=1}^{n-1}(100(x_i^2-x_{i+1})^2+(x_i-1)^2),\ x_i\in[-5,5] \qquad \text{Eq. 4}$$

This function is a minimization problem with a single global optimum at $[1]^n$. It has shown to be very difficult for many EAs and EDAs. An in-depth analysis of the landscape structure is given below to reveal some important information that determines the performance of EDAs.

This function can be regarded as the sum of two sub-functions:

$$f_{Rosen1}(X)=\sum_{i=1}^{n-1}100\cdot(x_i^2-x_{i+1})^2 \qquad \text{Eq. 5}$$

$$f_{Rosen2}(X)=\sum_{i=1}^{n-1}(x_i-1)^2 \qquad \text{Eq. 6}$$

It is well known see that f_{Rosen1} has numerous global optima in the bottom of a deep valley, which can be represented by:

$$X^{global}=[\alpha,...,\alpha^{2^{(i-1)}},...,\alpha^{2^{(n-1)}}] \qquad x_i\in[-5,5] \qquad \text{Eq. 7}$$

Note that since X is bounded in [-5, 5] in each dimension, the value of α should normally be between -1 and 1 because otherwise other elements in the array can easily exceed the boundary. Function f_{Rosen2} is a simple bowl-shaped landscape with $[1]^n$ as its global optimum, which is also one of the optima of f_{Rosen1} and thus the global optimum of the Rosenbrock function. Since there is no dependence, it can be easily solved by either EAs or EDAs.

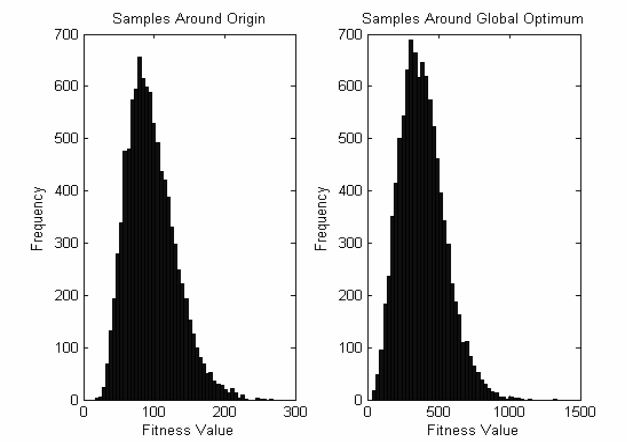


Figure 1. The distributions of fitness values in the 10D f_{Rosen}.

It is easy to see that f_{Rosen1} decides the global landscape structure due to the large factor 100. Furthermore, in some preliminary experiments, we found that the EDA could often converge to the origin, which is a global optimum of f_{Rosen1}.

The reason is that X^{global} is an array converging towards 0 when α is between -1 and 1 and even if α is not very close to 0, subsequent elements in the array may still quickly converge to 0. Furthermore, there is a global tendency that individuals close to the origin are likely to be better than those away from the origin because as long as each element in an individual is small, there is a good chance that f_{Rosen1} will produce a small value too, even if Eq. 7 is not satisfied. In other words, the density of good individuals around the origin is higher than other areas.

To demonstrate this point, 10,000 individuals were randomly generated within a 10D hyper-cubic area centred at the origin (i.e., [-0.5, 0.5] in each dimension). The distribution of fitness values of these individuals is plotted in Figure 1 (left) in which the majority of the individuals had fitness around 100. The same experiment was conducted with sampling centre at the global optimum (i.e., [0.5, 1.5] in each dimension). As a contrast, the fitness values of individuals shown in Figure 1 (right) were significantly worse.

Based on the above analysis, we can explain why this problem is difficult for EDAs. Although EDAs can easily find the global optima of f_{Rosen1} and f_{Rosen2} separately, a simple combination of these two functions creates much trouble. Due to the existence of the major attractor in the origin and the global structure, many promising individuals in the early stage of searching will be close to the origin and there is a high probability that EDAs will converge to it, instead of $[1]^n$, which is the true global optimum of f_{Rosen}. In other words, the overall structure identified by EDAs is different from the local structure around the global optimum and thus may mislead EDAs towards a non-optimal solution.

This issue was previously approached by incorporating clustering into EDAs in the hope of maintaining the global population diversity[3]. The general idea is to, in each generation, separate the population into several sub-populations and conduct independent searching in parallel. By doing so, EDAs are expected to be able to keep searching and maintaining different areas. However, due to the shape of the landscape, even multiple Gaussians may not be able to approximate it well enough and experimental results are not satisfactory[5].

Let's consider another question: what would happen after EDAs converge to the origin? Because f_{Rosen} is a unimodal function and the global optimum is connected to the origin via a valley, is it possible for EDAs to walk down the valley towards the global optimum, like other gradient-based algorithms? Even if it is hard to prevent an EDA from converging to the origin due to the misleading information that it encountered in the early stage of evolution, does it have any chance to get out of this trap? In order to do so, EDAs must be able to continuously sample individuals from the valley towards the global optimum. Since these individuals have better fitness than those around the origin, they will replace those old individuals and gradually shift the Gaussian away from the origin.

Unfortunately, when EDAs converge towards the origin, the size of their current search space may quickly shrink because elements in the covariance matrix may be getting close to 0, which is due to the similarity of promising individuals. This means that it is necessary to explicitly maintain the diversity of the probability distribution to prevent the variances from dropping too quickly so that EDAs may still have the power to keep searching.

3.2 The Sumcan2 Function

In order to demonstrate some issues of the existing way of applying clustering in EDAs, a new benchmark problem called Sumcan2 is proposed, which is partially based on the Summation Cancellation function [1]:

$$f_{Sumcan2} = \max\{f_{Sumcan}, f_{Local}\} \quad x_i \in [-5, 5] \quad \text{Eq. 8}$$

$$f_{Sumcan}(X) = \frac{1}{0.00001 + \sum_{i=1}^{n} |y_i|} \quad \text{Eq. 9}$$

$$\textit{where} \quad y_1 = x_1 - 3; \qquad y_i = y_{i-1} + x_i - 3$$

$$f_{Local}(X) = 5 \cdot \left(\prod_{i=1}^{n} \exp\left(\frac{-(x_i+3)^2}{4}\right) \right)^{1/n} \quad \text{Eq. 10}$$

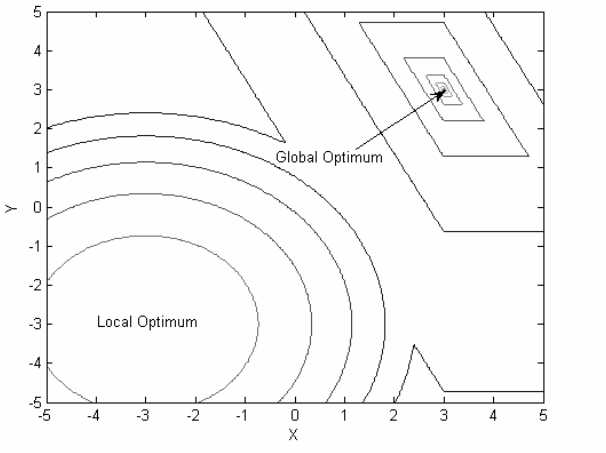


Figure 2. The contour of the 2D f_{Sumcan2}.

The contour of the 2D f_{Sumcan2} is plotted in Figure 2. In general, it has one global optimum created by f_{Sumcan} at $[3]^n$ with value 10^5 and one local optimum created by f_{Local} at $[-3]^n$ with value 5. The

reason for designing this function is to make it very difficult for most EAs and EDAs in order to highlight the performance of EDAs with clustering. This function is difficult for EAs in that the global optimum is the Summation Cancellation function, which contains strong dependences among variables and cannot be easily solved by algorithms without structure learning. It is also difficult for EDAs based on the one Gaussian model because it is multimodal and the local optimum is distant from the global optimum. Furthermore, f_{Local} has a relatively flat bell shape while f_{Sumcan} is like a very sharp spark, which means that random sampling is unlikely to generate many individuals close enough to the global optimum to have high fitness to be selected, compared to individuals around the local optimum. As a result, it is very likely that most promising individuals are initially from the basin of f_{Local} and consequently EDAs may be misled away from the global optimum and incorrectly converge to the local one.

It should be pointed out that this function also poses some difficulty for EDAs using clustering techniques in the traditional way because at the beginning of evolution, it is very likely that those selected individuals are distributed around the local optimum and none of them represents the area that contains the global optimum. Under this situation, no clustering techniques or statistical models could reveal the underlying structure of the problem and EDAs are likely to be failed.

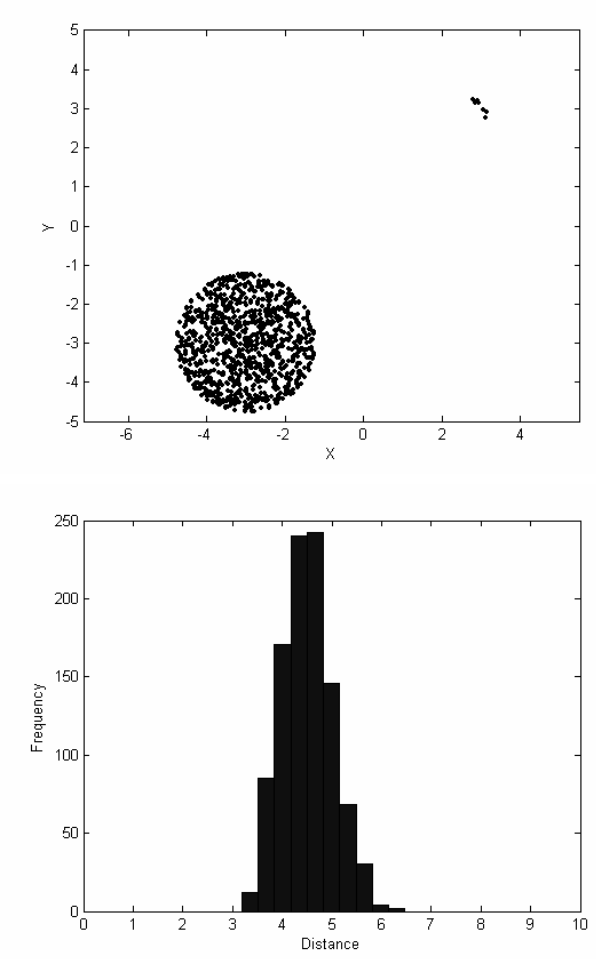


Figure 3. Sampling on the 2D (top) and 10 D (bottom) f_{Sumcan2}.

To have a better understanding of the situation, 10,000 individuals were randomly sampled in the search space with dimension equal to 2. After evaluation, the top 10% of individuals were selected and plotted in Figure 3 (top). Obviously, the majority of selected individuals were from the local optimum while only very few individuals represented the global optimum. No matter what kind of clustering algorithm is in use, it is very likely that only one cluster could be identified because those few points in the upper-right corner may be regarded as noise or outliers. In fact, when the sampling size is not very large, it is also possible that there is no individual from the global optimum due to some random factors. Note that although this problem has some similarity with the Rosenbrock function, there is no path connecting the two optima, which means that once EDAs get stuck at the local optimum, there is almost no chance to escape.

Furthermore, this situation could deteriorate as dimensionality goes up. The same sampling was conducted on the 10D Sumcan2 function and the distribution of distance (i.e., measured in terms of mean distance in each dimension) between those 10% selected individuals and the global optimum is plotted in Figure 3 (bottom). It is clear that all selected individuals were, on average, at least 3 units away from the global optimum in each dimension, which confirms that selected individuals are very unlikely to be around it. A larger experiment was conducted with 10 times the original population size (i.e., 100,000 individuals) and still no significant changes could be observed. This means that it is important to make sure that selected individuals do have a good coverage of the area that contains the global optimum before applying any clustering technique. Certainly, it is often impossible to check the condition in practice and simply applying clustering techniques could be of little help.

4. EXPERIMENTS

4.1 Simulations on the Rosenbrock Function

Experiments on f_{Rosen} were conducted with the following parameters: dimension=10, maximum number of generation=200, truncation selection=top 30% individuals. A few preliminary trials were run with population sizes from 200 to 2000. Unsurprisingly, this algorithm always got stuck at fitness around 7, which is similar to the results reported before[4, 5].

Figure 4 shows the evolution of the mean and standard deviation of each variable in a single trial with population size= 500. Since the global optimum of the 10D f_{Rosen} is at $[1]^{10}$, it may be expected that the means of variables should gradually move towards it. In Figure 4 (top), during the first few generations, the mean values were drifting around the origin because the origin is a major attractor based on the overall structure. After around 15 generations, mean values did start moving towards one, which is the correct direction. However, the algorithm quickly got stuck at somewhere in the valley between the origin and the global optimum after around 10 generations.

So what stopped this EDA from continuing its trip? The answer is in Figure 4 (bottom), which shows the standard deviation of each variable. It is clear that the standard deviations dropped very quickly and after about 25 generations, they were already around 10^{-3}. This means that the EDA was only searching an extremely limited area, like the tip of a needle. As shown in Figure 4 (top), at this stage, the EDA was on its way in the valley and still a bit far from the global optimum.

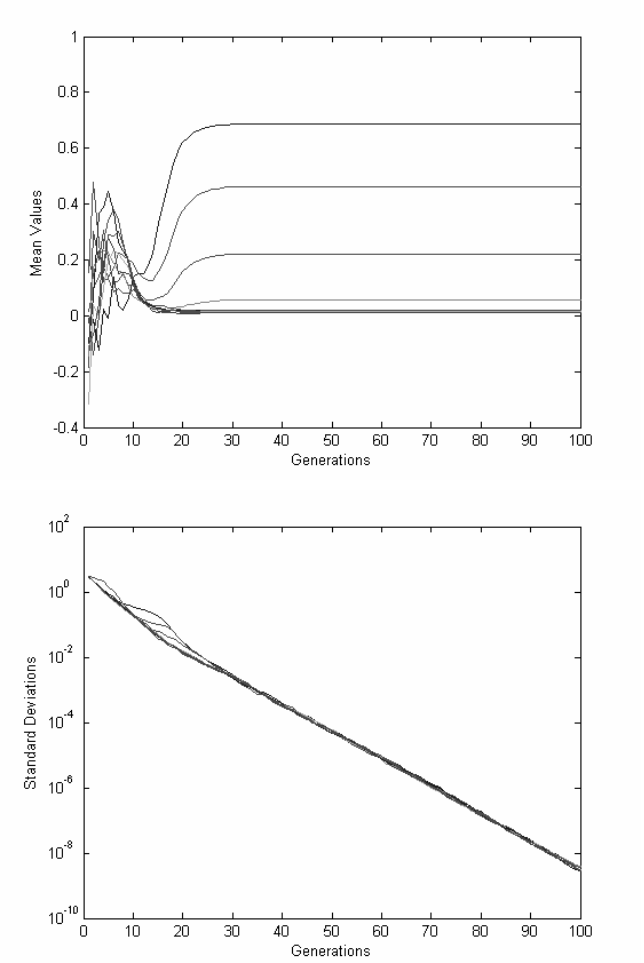


Figure 4. Convergence behaviour of the EDA on f_{Rosen} without diversity maintenance: mean values (top) and standard deviations (bottom).

Recall that in Eq. 3, Z is a matrix containing random numbers drawn from G (μ, δ^2) with μ=0 and δ=1. So, a simple way to maintain diversity is to use δ>1. The experimental results with population size=200 and δ=1.5 are plotted in Figure 5 from which it is clear that the performance of the EDA was dramatically improved. Figure 5 (top) shows (i.e., in a single trial) that the mean vector, after drifting around the origin for some generations, continuously moved towards the global optimum until the EDA successfully converged there. Figure 5 (middle) shows that the standard deviation values were orders of magnitude higher than before in the early stage and started quickly dropping after 140 generations when the mean vector was already very close to the global optimum, conducting fine local search. Figure 5 (bottom) shows the result averaged over 10 independent trials from which we can see that after 40,000 function evaluations, the best individuals found were often very close to 10^{-6} and still had the tendency to improve further. This result is comparable or better than most results previously reported without any additional computational cost. Certainly, keeping the diversity usually means sacrificing convergence speed for reliability. So, there is a tradeoff that needs to be taken into account when choosing the value of δ.

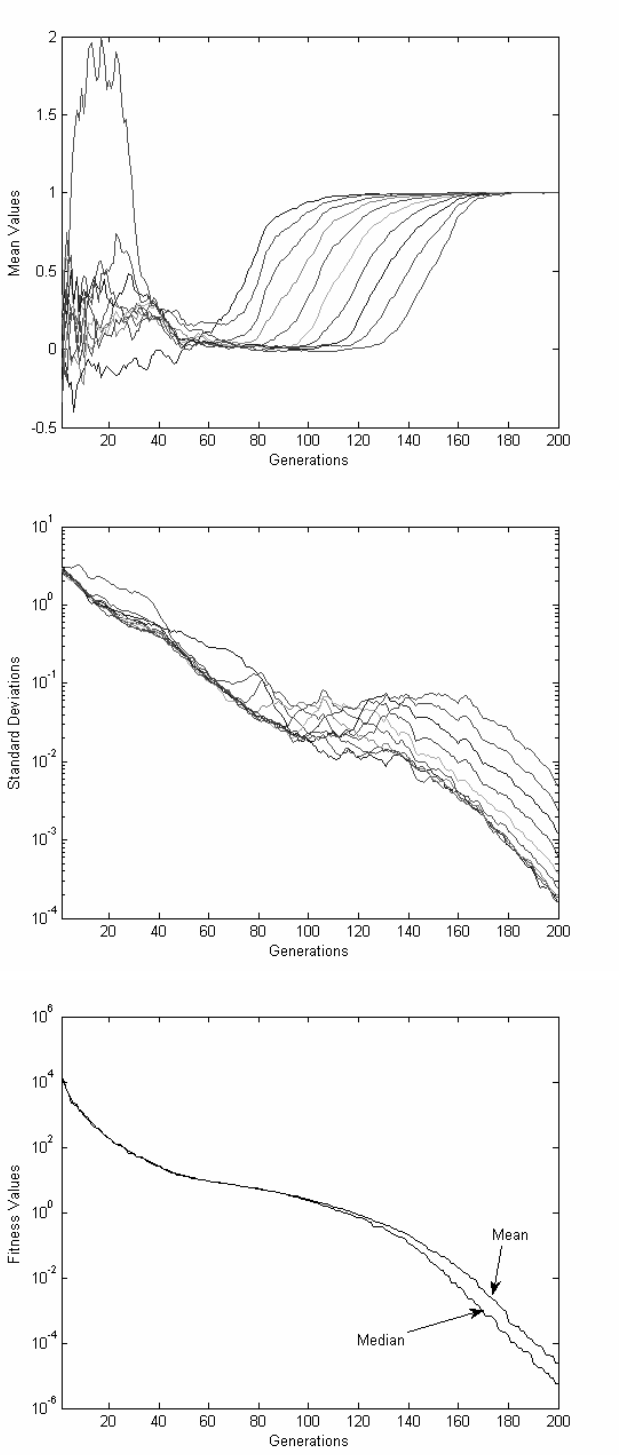


Figure 5. Convergence behaviour of the EDA on f_{Rosen} with diversity maintenance: mean values (top), standard deviations (middle) and fitness values vs. generation (bottom).

4.2 Simulations on the Sumcan2 Function

In order to demonstrate the difficulty of this problem, some experiments were first conducted on the 10D $f_{Sumcan2}$ with the standard EDA (population size=500, number of generations = 50).

Although the global optimum is at $[3]^{10}$, the EDA always quickly converged to $[-3]^{10}$ and got stuck at the local optimum. A single trial is picked up and the evolution of the mean values is plotted in Figure 6.

This means that without explicit diversity maintenance, the EDA is very likely to be misled by high quality local optima with relatively large basins. Since the EDA is assumed to be able to maintain only one Gaussian during evolution, it seems that clustering is a straightforward method to improve its performance. Traditionally, clustering is applied in each generation (Table 2). However, as shown in Section 3.2, it would be inappropriate to apply clustering techniques from the beginning of evolution in this case due to the high risk that the global optimum may not be included in the area to be clustered.

In order to address this issue, a new clustering scheme is proposed in which an EA with diversity maintenance ability is run first for some generations and some clustering technique is then applied on selected individuals from the final population. Finally, EDAs continue the searching by running on each cluster separately (Table 2). The basic idea is to use this EA to do some rough searching first. Although it may not be able to find individuals of very high quality, it may still gradually cluster individuals in promising areas. As a result, it is more likely that the global optimum is within one of the clusters of individuals than in the traditional framework.

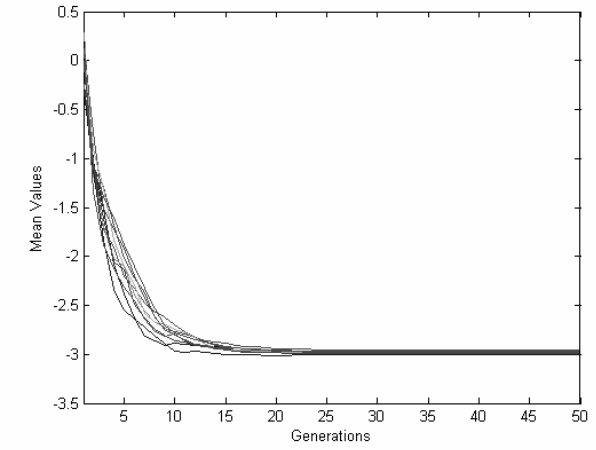


Figure 6. Convergence behaviour of the EDA on $f_{Sumcan2}$.

Table 2. Two schemes for EDAs with clustering.

Traditional Framework	New Framework
• In each generation, cluster selected indiviuduals into k clusters. • Estimate the probability distriubtion P_i (i=1,...,k) of each cluster of individuals. • Create a new population by sampling from each P_i seperately.	• Conduct searching by an EA with diversity maintenance ability. • Cluster selected individuals from the final population into k clusters. • Run the EDA k times using different clusters as the initial population.

A simple (μ+λ) ES (Evolution Strategy)[15] was used to conduct the rough searching in which new individuals are generated though Gaussian mutation with fixed diagonal covariance matrix. Unlike traditional ESs, in this ES, newly generated individuals could only replace their corresponding parents provided that they have better fitness values, which is very similar to the idea of deterministic crowding [10].

In the first step of the experiments, the parameters of the ES were chosen as: population size=1000, standard deviation=0.5, number of generations=200. After $2x10^5$ fitness evaluations, top 30% individuals were selected to be clustered. The distribution of those 300 individuals is plotted in Figure 7 in which each vertical line represents a certain axis (dimension) and each individual is represented by a single curved line crossing all vertical lines whose intersection with each vertical line is determined by its value at the corresponding dimension. By doing so, the clustering patterns of data in high dimensional spaces could be intuitively observed. It is easy to see that, after the preliminary searching by the ES, individuals did present some clear clustering patterns, which reflected the underlying problem structure.

Next, since there were clearly two clusters as shown in Figure 7, the k-mean algorithm (k=2) was used to do clustering. Figure 8 shows the box whisker plots of those 300 selected individuals grouped into two clusters. Since the median values of cluster 1 were close to $[3]^{10}$ and the median values of cluster 2 were close to $[-3]^{10}$, we can see that these two clusters of individuals roughly corresponded to those two optima.

The effectiveness of clustering is shown by the Silhouette plot and the slice plot of the first two dimensions in Figure 9. The Silhouette plot shows the distance among points in their own cluster compared to the distance to points in other clusters. As shown in Figure 9 (top), the Silhouette values of all points were more than 0.9 showing that points (individuals) in those two clusters were well separated from each other. Furthermore, in Figure 9 (bottom), there were clearly two clusters of individuals corresponding to the local optimum and global optimum respectively (See Figure 3 for comparison), which provided a good starting position for EDAs to do further searching. The slice plots of other pairs of dimensions also had this kind of pattern.

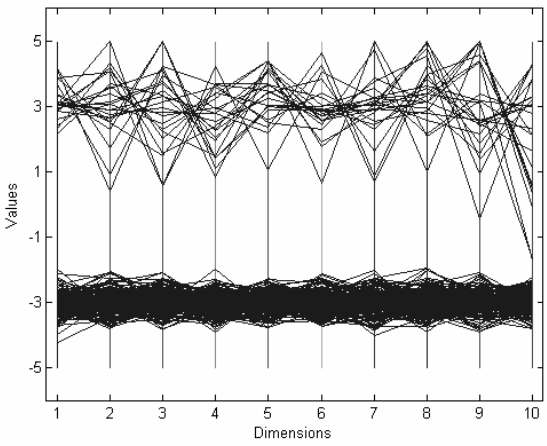


Figure 7. The clustering pattern of selected individuals in the final population of the ES.

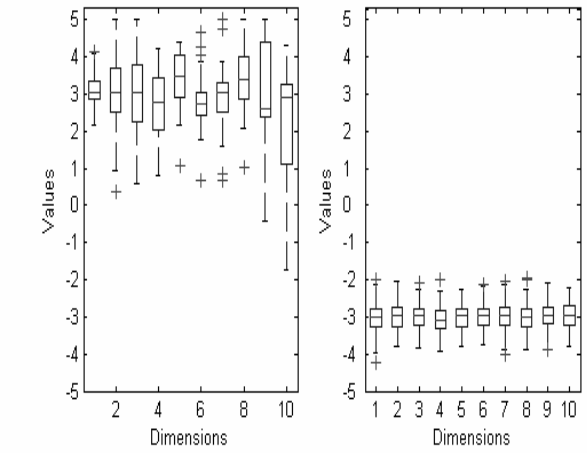
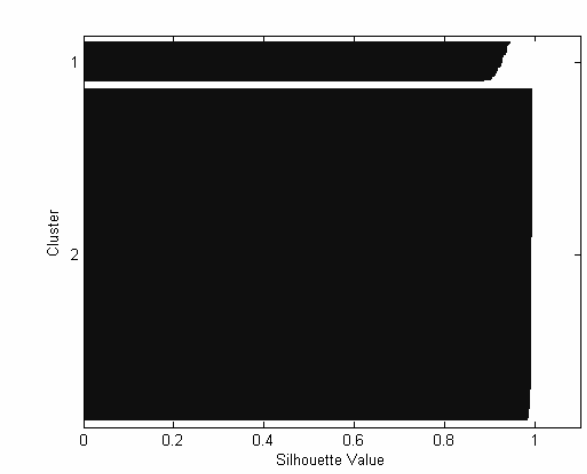


Figure 8. Distribution of individuals: Cluster 1 (left) and Cluster 2 (right).



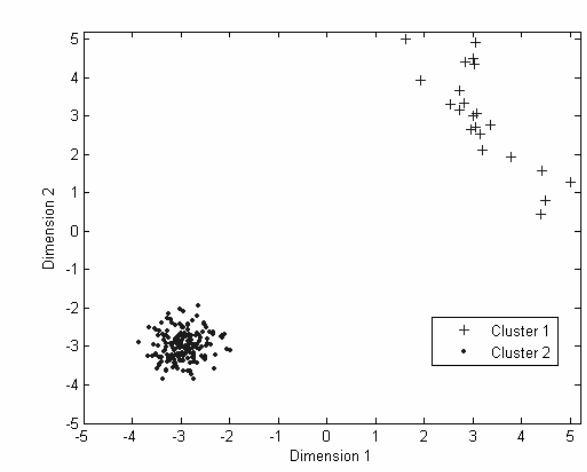


Figure 9. The effectiveness of clustering: Silhouette plot (top) and 2D slice plot (bottom).

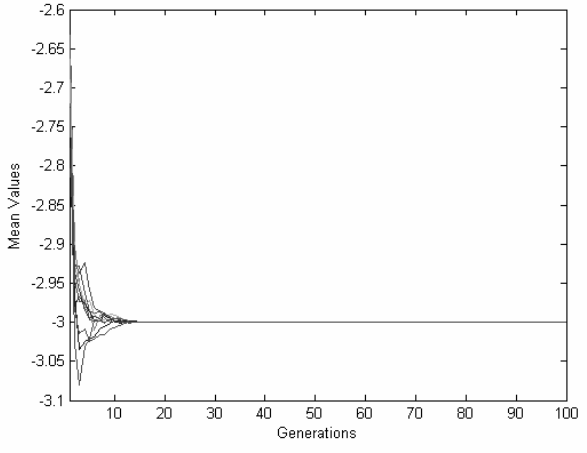
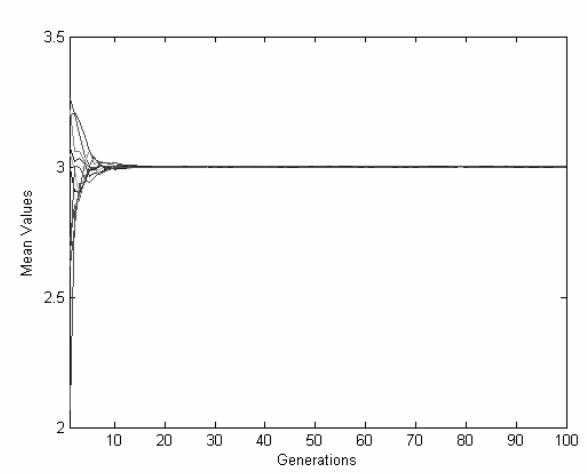


Figure 10. Convergence behaviour of the EDA on f_{Sumcan2} with cluster 2 as the initial population.



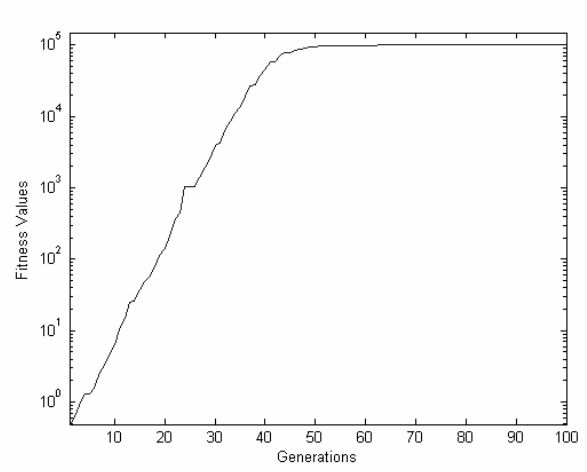


Figure 11. Convergence behaviour of the EDA on f_{Sumcan2} with cluster 1 as the initial population: mean values (top) and fitness vs. generation (bottom).

Finally, the EDA was run twice with each time using a different cluster of individuals as the starting point and the results in a single trial are plotted in Figures 10 & 11 respectively. Figure 10 shows that when using cluster 2 as the starting point, the EDA got stuck at the local optimum at $[-3]^{10}$ again. As a contrast, the EDA using cluster 1 as the starting point quickly converged to $[3]^{10}$ and found the global optimum with value 10^5.

In the above experiments, we demonstrated how a difficult multimodal problem was solved with the help of a clustering algorithm and an ES, which prepared individuals in a good format before undertaking clustering. This approach is different from previous work in that those selected individuals are generated by another algorithm and only need to be clustered once. This is particular useful when randomly generated individuals are less likely to reveal the important structure of the problem to be solved due to the small basin size of the global optimum and/or the existence of a number of good local optima with large basins.

Another potential issue of the traditional method is that although EDAs work on each cluster of individuals separately, newly generated individuals are combined together and some are selected to undertake further clustering. Due to the lack of explicit diversity maintenance mechanism, clusters corresponding to good local optima with large basins may grab more and more individuals from small clusters because the number of individuals generated is usually in proportion to the size of the cluster/average fitness of individuals and good individuals are more likely from those currently promising clusters.

There are some practical issues that need consideration in the proposed approach. For example, the type of EA used to do rough searching and its parameters may have some influence on the distribution of individuals in the final population, which directly decides the performance of clustering algorithms. Although the performance of the ES was satisfactory and relatively robust in our experiments, this may not always be true in other situations. It may also make sense to design EAs that are not intended to find individuals that are as good as possible but to identify and maintain promising areas, which may contain the global optimum.

5. CONCLUSION

This paper is dedicated to an in-depth analysis of the structure of two continuous problems in order to understand what makes them difficult for EDAs based on the one Gaussian model. Through a set of experiments, we explored the evolution process of an EDA to have a deep insight into how its performance was influenced by the properties of each problem. We pointed out that diversity maintenance plays a key role in the success of EDAs. It has shown that, by keeping the variances of variables from quickly dropping to zero, the performance of the EDA on the Rosenbrock function was dramatically improved. In order to handle multimodal problems, clustering was incorporated into EDAs and a novel three-step-scheme containing EAs, clustering algorithms and EDAs was proposed, which may overcome some difficulty faced by traditional methods. Note that although experiments were conducted with an EDA based on Cholesky decomposition, these ideas could easily be transferred into other EDAs. Certainly there are also some issues with these proposed methods that must be carefully investigated before they can be successfully applied in other problems, which will be pursued in the future.

6. ACKNOWLEDGEMENT

This work was supported by an Australian Postgraduate Award granted to Bo Yuan.

7. REFERENCES

[1] Baluja, S. and Davies, S. Using Optimal Dependency-Trees for Combinatorial Optimization: Learning the Structure of the Search Space. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, 30-38.

[2] Bosman, P.A.N. and Thierens, D. *An Algorithmic Framework For Density Estimation Based Evolutionary Algorithms*. Technical Report UU-CS-1999-46, Utrecht University, 1999.

[3] Bosman, P.A.N. and Thierens, D. *Mixed IDEAs*. Technical Report UU-CS-2000-45, Utrecht University, 2000.

[4] Cho, D.-Y. and Zhang, B.-T. Evolutionary Continuous Optimization by Distribution Estimation with Variational Bayesian Independent Component Analyzers Mixture Model. In *Proceedings of Parallel Problem Solving from Nature VIII*, 2004, 212-221.

[5] Kern, S., Müller, S.D., Hansen, N., Büche, D., Ocenasek, J. and Koumoutsakos, P. Learning Probability Distributions in Continuous Evolutionary Algorithms - A Comparative Review. *Natural Computing, 3*, 1 (2004), 72-112.

[6] Kullback, S. and Leibler, R.A. On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 1 (1951), 79-86.

[7] Larrañaga, P., Etxeberria, R., Lozano, J.A. and Pena, J.M. *Optimization by learning and simulation of Bayesian and Gaussian networks*. Technical Report EHU-kZAA-IK-4/99, University of the Basque Country, 1999.

[8] Larrañaga, P. and Lozano, J.A. (eds.) *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, 2001.

[9] Lu, Q. and Yao, X. Clustering and Learning Gaussian Distribution for Continuous Optimization. *Accepted by IEEE Transactions on Systems, Man, and Cybernetics, Part C* (2004)

[10] Mahfoud, S.M. Crowding and Preselection Revisited. In *Proceedings of Parallel Problem Solving From Nature II*, 1992, 27-36.

[11] Mühlenbein, H. and Paaß, G. From Recombination of Genes to the Estimation of Distributions: I. Binary Parameters. In *Proceedings of Parallel Problem Solving from Nature IV*, 1996, 178-187.

[12] Nash, J.C. *Compact numerical methods for computers: linear algebra and function minimisation*. Adam Hilger, 1990.

[13] Paul, T.K. and Iba, H. Real-Coded Estimation of Distribution Algorithm. In *Proceedings of The Fifth Metaheuristics International Conference*, 2003.

[14] Ripley, B.D. *Stochastic Simulation*. John Wiley & Sons, 1987.

[15] Schwefel, H.-P. *Evolution and Optimum Seeking*. Wiley, New York, 1995.