Large scale continuous EDA using mutual information

Xu, Qi; Sanyang, Momodou; Kaban, Ata

DOI:
10.1109/CEC.2016.7744260

Document Version
Peer reviewed version

Large Scale Continuous EDA Using Mutual Information

Qi Xu
School of Computer Science
University of Birmingham
Email: qxx506@student.bham.ac.uk

Momodou L. Sanyang
School of Computer Science
University of Birmingham
Email: m.l.sanyang@cs.bham.ac.uk

Ata Kabán
School of Computer Science
University of Birmingham
Email: a.kaban@cs.bham.ac.uk

***Abstract*—Most studies of Estimation of Distribution Algorithms (EDA) are restricted to low dimensional problems due to EDA being susceptible to the curse of dimensionality. Among methods that try to scale up EDA to high dimensional problems, EDA-MCC was recently proposed. It controls the complexity of the search distribution by thresholding correlation estimates as a means to approximate the prominent dependency structure among the search variables and discard irrelevant detail.**

However, it is known that the correlation coefficient can only determine statistical dependence when the data distribution is Gaussian. In this paper, we develop a new variant of EDA-MCC called EDA-MCC-MI which uses mutual information (MI) estimates to determine dependencies between the search variables, replacing linear correlation. Our method is in a better position to determine the correct dependency structure than the EDA-MCC can do, simply because MI is zero if and only if the variables are independent, whereas a zero correlation does not imply independence in general. Empirical comparison results show that EDA-MCC-MI is never worse than EDA-MCC even when the search distribution is Gaussian. Our implementation employs a nonparametric MI estimator, hence it is easily extensible to any other, non-Gaussian search distribution.

I. Introduction

Estimation of distribution algorithms (EDA) are a relatively new variety in terms of evolutionary algorithms (EA) [1]. EDA guides the search for optimum by creating new candidate solutions based on the probabilistic model it builds on the most fit individuals. In comparison with traditional EAs, EDA generates a model which obeys a specified probability distribution to extract and utilise the structure of the current best individuals, rather than taking it into no account [2].

Much research into EDAs has been done in the past. Among the earliest proposed EDAs are Population Based Incremental Learning (PBIL) [3] and UMDA$_c^G$ [1] adopt univariate Gaussians. That means, all the input variables are assumed to be independent from each other. Due to this assumption, these methods are not able to solve problems in which the variables have strong dependencies, even though the algorithms are fast and easy to implement. For better performance, other EDAs have been proposed, such as EMNA and ENGA [1]. Both of these use maximum likelihood estimation (MLE), to estimate a full multivariate model in EDA. These algorithms are based on multivariate Gaussian model, therefore they are able to find out and make use of more dependencies among variables. Later, a new algorithm EEDA was proposed [4]. This algorithm scales up EMNA and ENGA by decomposing the covariance matrix of the Gaussian model. Even though Gaussian-based EDAs are most widely employed, EDAs based on other distributions are also studied. For example, in [5], and references therein, multivariate heavy tailed distributions were investigated.

A large amount of research focuses on low and moderate dimensional problems (such as problems of size lower than 100-D), since EDAs are highly effective in such cases. However, high dimensional problems often lead to poor performance. This is because EDAs suffer from the *curse of dimensionality* [6]. Since the new solutions generated by EDAs are completely based on their probabilistic model estimation, it is inevitable to suffer from this curse. The reason is that, in order to obtain an accurate model, there are mainly two ways: add constraints to the model by exploiting some prior knowledge about the problem, or build the model with a large population [6]. The first approach is problematic when there is no prior knowledge, and EDA must estimate a model only based on the input data. The second approach is often feasible in low dimensional cases, but when it comes to high dimensional problems, the computational cost grows dramatically. Taking a deeper insight into this problem, when the high fitness individuals are selected from the population, they should form a large enough set of points in order to produce accurate estimates about the global search space. Therefore, the population has to be very large, as the problem dimension grows. In univariate models, this is not necessarily observed, because the problem is approached in each dimension separately, thus if the population is large enough for one dimension, the model may be accurately estimated. However, that is a simplified inaccurate model with limited abilities. When it comes to multivariate models, since the dependencies of variables should be considered, a large population must be involved for the accuracy of the model estimation, which results in a high computational complexity. One can employ randomised dimensionality reduction for a cheap and effective way to decrease the dimension of the search space [7], [8]. Other approaches include the use of multiple populations to overcome the above mentioned problems [9].

EDA with model complexity control (EDA-MCC) is a recent approach [10] which models the selected high fitness individuals by a constrained multivariate Gaussian. Because the model complexity is constrained, it has a significantly

lower computational cost than many traditional EDAs, and it is less likely to overfit. It has been demonstrated to give more accurate results than most univariate models based methods and unconstrained multivariate model based ones. It somewhat relates to the idea of factorising the multivariate distribution using a Gaussian network of [11] – the differences between these approaches are explained in [10].

EDA-MCC uses the thresholded linear correlation coefficient estimates to control the model complexity. Small correlations are discarded, and any search variables that only have small correlations are treated as independent. It is known, however, that the linear correlation coefficient is only able to represent linear dependencies, and this is only sufficient when the variables are Gaussian. However, in EDA, the model is built on the *selected* individuals. After selection, the variables may no longer be Gaussian in general, even if the sampling distribution was Gaussian. Figure 1 shows an example. A zero correlation does not imply independence unless if the variables are Gaussian. Moreover, in some cases onr might like to use a non-Gaussian search distribution. Then, even the sampled new generation will not be Gaussian distributed. Our method proposed in this paper is applicable in principle to all such cases.

In this paper, we propose an alternative technique of determining dependencies between variables in EDA-MCC – the mutual information (MI) based method, which replaces linear correlation coefficient estimation with MI estimation in EDA-MCC. We call our new method EDA-MCC-MI. The reason is that, contrary to correlation, the MI is only zero if the variables are independent. Experiments on five well-known benchmark functions show that EDA-MCC-MI has the ability to outperform the original EDA-MCC.

II. The method of EDA-MCC

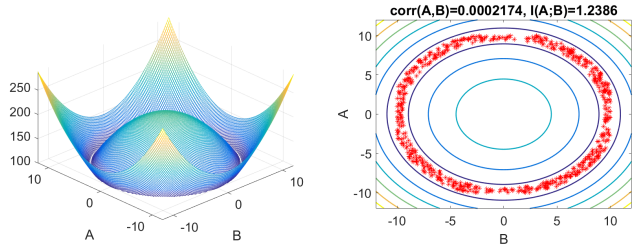
EDA-MCC [10] is composed of two main steps: 1) Identification of Weakly Dependent variables (WI) is done by collecting together all those search variables whose correlations with other search variables are all below a pre-specified threshold. 2) All the rest of the search variables are used in the Subspace Modelling (SM) step.

The weakly dependent variables are then modelled by univariate Gaussian distributions, while the rest of the variables are randomly grouped in equal size non-overlapping groups, and each group is modelled by a multivariate Gaussian. This algorithm has been demonstrated to outperform many traditional EDAs because the complexity of the model is controlled by the decomposition of the problem into the WI and SM steps.

Using this strategy, in comparison with EMNA, the computational cost is reduced, but the accuracy of modelling is not heavily affected.

III. Our approach: EDA-MCC-MI

Figure 1 shows the difference between correlation and MI for two variables A and B which have a non-linear correlation. The linear correlation coefficient is close to zero in this case,



(a) An example function which has optima on a circle (b) Points near optima after selection

Fig. 1: The difference between correlation and mutual information when the variables have a non-linear correlation.

which means that the variables are close to being uncorrelated. But they are not independent, and indeed the MI equals 1.2386 – which shows that the variables are in fact strongly dependent. In other words, linear correlation is unable to show the independence between variables when the data distribution is non-Gaussian.

A. Mutual Information in WI

1) Theoretical Definition: In information theory, mutual information is defined as follows [12]. First, the information entropy of a random variable A is defined as:

$$H(A) = E[I_c(A)], \tag{1}$$

where E is the expected value operator, and $I_c(A)$ denotes the self-information of A. $I_c(A)$ is defined as:

$$I_c(A) = -\ln p(A), \tag{2}$$

where $p(A)$ is the probability density function (PDF), and the logarithm is in base 2 (so the amount of information is measured in bits).

For two random variables A and B, the conditional entropy of A given B is defined as:

$$H(A|B) = \sum_A \sum_B p(A,B) \ln \frac{p(B)}{p(A,B)}. \tag{3}$$

where the notation assumes discrete valued variables. In the continuous case the sums become integrals.

The mutual information $I(A;B)$ is then defined as:

$$I(A;B) = H(A) - H(A|B) = H(B) - H(B|A). \tag{4}$$

For illustration, Figure 2 is a Venn diagram which shows the relationships of entropy, conditional entropy and mutual information. Intuitively, the mutual information quantifies the amount of information (in bits) that two random variables have about each other.

From Figure 2, it is seen that, the area contained by both circles is the joint entropy, i.e. the intersection of $H(A)$ and $H(B)$. This part shows the dependency between A and B. If $I(A;B) = 0$, which means the circles are not overlapped, then the two variables are independent. Moreover, the converse is also true: A and B are independent only if $I(A;B) = 0$.

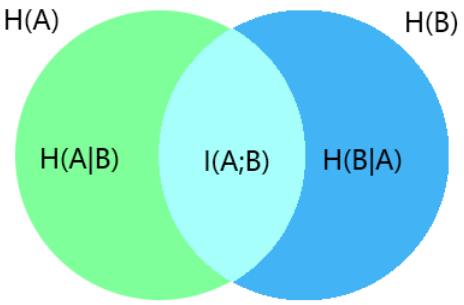


Fig. 2: Mutual information

An alternative way to write $I(A;B)$ is:

$$I(A;B) = \sum_A \sum_B p(A,B) \ln \frac{p(A,B)}{p(A)p(B)} \tag{5}$$

which is the Kullback-Leibler distance between the joint PDF and the product of the two marginal PDFs.

The larger $I(A;B)$ is, the more strongly the variables are dependent. Theoretically, the mutual information can be infinity for two infinite variables, because if A and B are dependent, then the larger the population size is, the more information they will share. However, in finite case, especially when the population size is fixed, the value of mutual information can be restricted in a specified range, and then the dependency between two variables can be measured.

2) Estimation of MI: Adaptive Histogram Method: In our approach we will use the MI estimator of [13], which is an adaptive histogram method. We have chosen this because of its efficiency and generality. Since it is non-parametric, it makes our approach applicable in principle to any search distribution in EDA.

This MI estimation algorithm divides the set of points into a finite number of disjoint rectangular partitions, and the probability density is estimated in each partition simply by counting. This MI estimate is computed as follows:

$$\hat{I}(A;B) = \sum_{k=1}^{m} \frac{N_k}{N} \ln \frac{N_k/N}{(N_{a,k}/N)(N_{b,k}/N)}, \tag{6}$$

where N is the number of points, m is the number of partitions, N_k is the number of points in the kth partition, and $N_{a,k}, N_{b,k}$ are the number of points whose first / second variable falls within the limits of partition k. This is indeed a sample estimate of the theoretical value of the MI in the form given in eq.(5). More analysis about this estimator may be found in [13], and an empirical comparison in [14] found it most efficient and accurate among other estimators tested.

Now we give a summary of the partitioning procedure. For more details see [13], [14]. It is a recursive method, which starts from a single partition defined as the smallest rectangle that contains all N points. Provided that the partition contains at least 2 points, it then attempts to split up the partition into 4 equally sized disjoint rectangles by halving both horizontally and vertically. A χ^2 goodness of fit test is applied to decide whether to keep the split or not. If the split passes this test, then the procedure continues recursively in each of the newly created four partitions. The procedure terminates when no partition can be split further.

Once the partitions are created, then the probability density is estimated simply by counting.

B. Improved EDA-MCC: EDA-MCC-MI

Our proposed algorithm, EDA-MCC-MI, is a direct extension of EDA-MCC proposed by [10]. It consists of replacing the computation of linear correlation coefficient estimates by computation of mutual information estimates. Algorithm 1 gives a schematic pseudo-code of EDA-MCC-MI.

Function *EDA-MCC-MI*
Input: Dimension D, Population size M, M_{sel}, m_c, threshold θ
Output: $\mathcal{P}$
Initialise a population $\mathcal{P}$ of M individuals
while *stopping criterion is not met* **do**
  Estimate fitness of all individuals
  Select $M_{sel} \leq M$ individuals from $\mathcal{P}$
  $X \leftarrow$Ramdomly sample $m_c \leq M_{sel}$ individuals from M_{sel} selected ones

  WI step:
  Compute the MI matrix $\mathcal{I}$ from X, $\mathcal{I}_{ij} = MI(X_i, X_j), i,j \in [1,D]$
  $\mathcal{W} \leftarrow \{X_i | \mathcal{I}_{ij} < \theta, \forall j \neq i\}$
  for $k = 1 : |\mathcal{W}|$ **do**
    Estimate univariate model of variable k for the M_{sel} selected individuals.
  end

  SM step:
  $\mathcal{S} \leftarrow [1,D] - \mathcal{W}$.
  Randomly divide $\mathcal{S}$ into $\lceil |\mathcal{S}|/c \rceil$ non-overlapping subsets.
  for $k = 1 : \lceil |\mathcal{S}|/c \rceil$ **do**
    Estimate a multivariate model for the M_{sel} selected individuals.
  end

  $\mathcal{P}' \leftarrow$ new individuals sampled based on the two models respectively.
  $\mathcal{P} \leftarrow \mathcal{P}'$.
end

Algorithm 1: EDA-MCC-MI

IV. Experimental Studies

A. Experimental Setup

The goal of our experiments is to establish whether mutual information is a viable, and potentially more advantageous alternative to the correlation coefficient in the context of EDA search. We conduct experiments comparatively with two algorithms: EDA-MCC and our modified version EDA-MCC-MI. As [10] mentioned, since EDA-MCC applies UMDA$_c^G$ [1]

in WI process and EEDA [4] in SM process, and it outperforms both of the algorithms in many cases, UMDA$_c^G$ and EEDA are not considered in our experiments.

We implemented our new algorithm in a combination of multi-threaded C++ and Matlab for computational efficiency[1].

1) Test Functions: Table I lists the test functions used in our experiments. They are selected from the benchmark functions used in the CEC'2005 Special Session. Definitions and descriptions of these functions can be found in [15]. We have chosen these five functions to cover different characteristics, and they belong to the following three groups:

1. Separable unimodal problems: F_1 and F_2.
2. Separable problems with many local optima: F_3 and F_4.
3. Non-separable problem: F_5.

2) Parameter Settings: Since the experiments are set up to compare the 2 algorithms, the parameters are similar with those set in [10]. Two problem sizes are tested in the moderate to high regime: $D = 100$ and $D = 500$. Three population sizes $M \in \{300, 1000, 2000\}$ are utilised for the 100D problems and $M = 1000$ for the 500D problems. The maximum fitness evaluation FE is set to $FE = 10000 \times D$. The maximum iteration is FE/M and the algorithm terminates when the iteration exceeds this value. For the experiments in this section, we set $m_c = 100$ and $\theta = 0.3$ for WI, because they are recommended in [10]. For SM, $c = \min\{D/5, M/15\}$ is used, for reasons expained in [16]. Some experimental analysis on the influence of θ will be given in Section V. All the experiments in this section were ran 25 times, and the reported results are the average values $\pm$ standard deviations of the best fitness value found in these 25 independent repetitions. The experiments in Section V were ran 10 times, and the same statistics are reported.

B. Experimental Results on 100-D Functions

Figure 3 shows the results on F_1–F_5, when the problem dimension is 100 and the population size is 300. We see that EDA-MCC-MI appears to outperform EDA-MCC in all the experiments, as it achieves lower fitness values in all minimisation problems tested. Table II shows the corresponding rank sum test results comparing EDA-MCC and EDA-MCC-MI. Any significant outperformance is shown in bold. We see that, even though EDA-MCC-MI appears better than EDA-MCC in all the plots in Figure 3, the statistical test detects no statistically significant differences in the performance of these two methods in most cases, and statistical significance is only detected in problem F_2.

Table III shows the statistical results between EDA-MCC and EDA-MCC-MI on 100D problems, when the population size is 1000 and 2000. It is seen that, the performance of the two algorithms is very close, and EDA-MCC-MI only shows significant outperformance on F_1 when $M = 2000$. Both the algorithms perform well on F_1 and F_4, and when $M = 2000$, they also achieve a result close to the optimum on F_2. The algorithms fail to find the global optimum on problems F_3 and F_5. The reason is that, when M grows, the estimation of the model becomes more accurate, and thus it is more likely for the algorithm to reach the global optimum.

Hence, so far we see that MI is indeed a viable alternative, and is never worse than correlation. One might hope for significant outperformance, however recall that here we are just using a Gaussian search distribution. Although MI is able to capture non-linear and higher order dependencies from the selected population (see e.g. Figure 1), the subsequent modelling of these selected individuals by a Gaussian has rather little ability to make use of the detected dependencies. However, the fact that EDA-MCC-MI turned out no worse than EDA-MCC even in this restricted setting, and was even slightly better in some cases is a strong evidence of its potential to be built in more flexible non-Gaussian search distributions for improved EDA search in future work.

C. Results on 500-D Problems

Our results on 500D problems are presented in Figure 4. We again see that EDA-MCC-MI achieves lower fitness values than EDA-MCC on all the 5 problems. In Sphere function, EDA-MCC has the average fitness value slightly lower than 1, however, EDA-MCC-MI has the value lower than 10^{-20}, which is very close to the global minimum of zero. In the Ackley function, EDA-MCC has the average fitness value slightly lower than 10^{-2}, and EDA-MCC-MI has the value lower than 10^{-12}. However, again the statistical analysis finds no significant differences in either of these two test problems. However, the statistical test did confirm that EDA-MCC-MI is significantly better than EDA-MCC on the remaining three problems, namely the shifted Elliptic, the shifted Rastrigin and the shifted Rosenbrock functions.

From the results on both 100 and 500 dimensional problems, we find that EDA-MCC-MI performs better than EDA-MCC in all the experiments, although the differences are not always statistically significant. Therefore EDA-MCC-MI shows remarkable effectiveness on both moderate and high dimensional problems tested, and on both separable and non-separable problems. It also performs well in both small population sizes and large population sizes. From analysing the effect of the threshold parameter and the proportion of weakly vs. strongly correlated variables, presented in the next section – it appears that the reason we find more significant outperformance in the case of our higher dimensional experiments is not due to detecting complicated dependencies but by contrary, due to less variables passing the threshold for multivariate modelling (since the estimated MI values are low when the number of points is small), which prevents inaccurate model building when the dimensionality grows while the population size remains low. This is also a useful feature to have in practice.

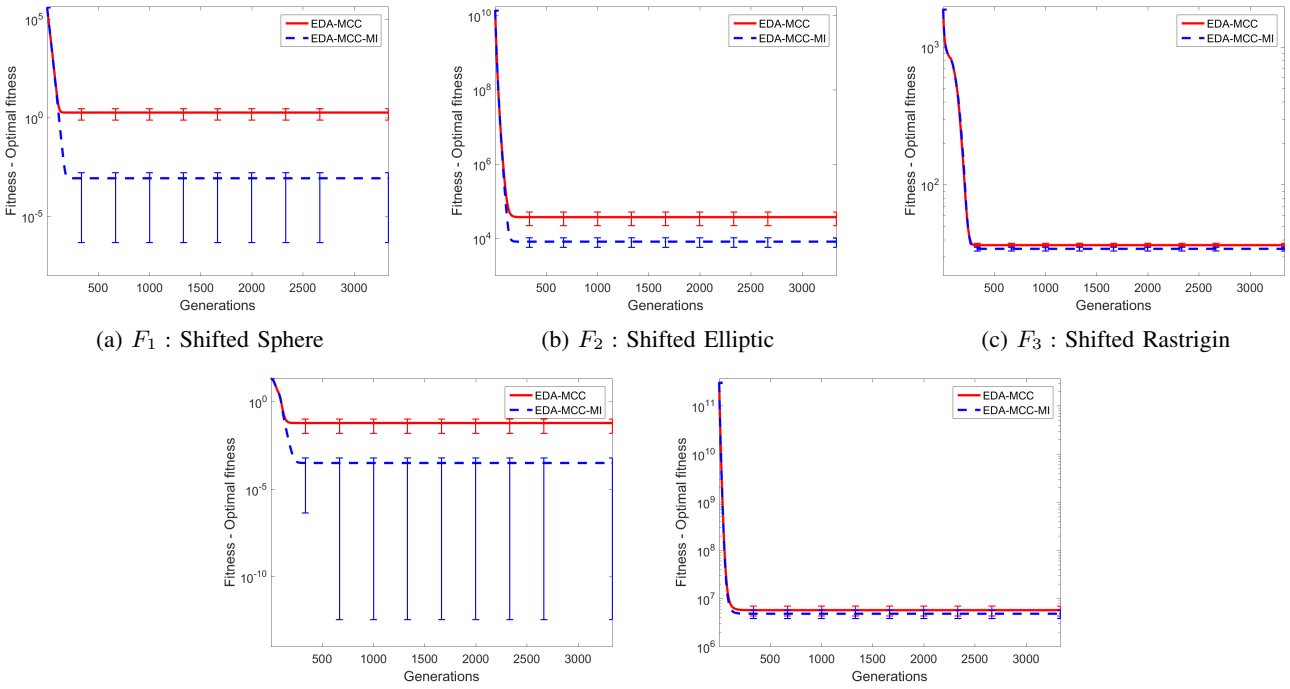
V. Influence of Parameter θ

In order to gain more insights into the working of EDA-MCC-MI, and to uncover reasons behind our experimental findings, we ran more experiments investigating the effect of the threshold θ in deciding the proportion of weakly dependent

[1]The code is freely available at: https://github.com/qxandy/edami

TABLE I: Benchmark functions used in our experiments

No.	Name	Expression
F_1	Shifted Sphere	$F(\mathbf{x}) = F_{sphere}(\mathbf{z})$, $F_{sphere}(\mathbf{x}) = \sum_{i=1}^{D} x_i^2$
F_2	Shifted Elliptic	$F(\mathbf{x}) = F_{elliptic}(\mathbf{z})$, $F_{elliptic}(\mathbf{x}) = \sum_{i=1}^{D} (10^6)^{\frac{i-1}{D-1}} x_i^2$
F_3	Shifted Rastrigin	$F(\mathbf{x}) = F_{rastrigin}(\mathbf{z})$, $F_{rastrigin}(\mathbf{x}) = \sum_{i=1}^{D} (x_i^2 - 10\cos(2\pi x_i) + 10)$
F_4	Shifted Ackley	$F(\mathbf{x}) = F_{ackley}(\mathbf{z})$, $F_{ackley}(\mathbf{x}) = -20exp(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}) - exp(\frac{1}{D}\sum_{i=1}^{D}(2\pi x_i)) + 20 + e$
F_5	Shifted Rosenbrock	$F(\mathbf{x}) = F_{rosenbrock}(\mathbf{z})$, $F_{rosenbrock}(\mathbf{x}) = \sum_{i=1}^{D-1} 100((x_i^2 - x_{i+1})^2 + (x_i - 1)^2)$



(a) F_1 : Shifted Sphere (b) F_2 : Shifted Elliptic (c) F_3 : Shifted Rastrigin

(d) F_4 : Shifted Ackley (e) F_5 : Shifted Rosenbrock

Fig. 3: Evolutionary curves for performance comparison between the use of MI and correlation coefficient in EDA-MCC for functions $F_1 - F_5$. $D = 100$, $M = 300$ and Budget size is $1 \times 10^4 \times D$

vs. strongly dependent variables. In these experiments, we varied $\theta \in [0.01, 0.09] \cup [0.1, 1.0]$. The experiments are run on all the 5 benchmark functions, with dimension $D = 100$ and population size $M = 1000$. For each function, the results reported are averages from 10 independent repetitions. Figures 5 – 7 show the results.

Recall that $F_1 - F_4$ are separable functions, and F_5 is a non-separable function. For separable functions, it is assumed that independent variables can obtain a good result, therefore we would expect that larger values of θ will work well. For the non-separable function, in turn, more dependencies are needed for a better result, so we expect that θ should be smaller for this purpose. However, this hypothesis assumes an accurate model estimation. When the population size is too small, or due to the bias introduced by excessive model complexity restriction, these other factors may also influence the effects of the threshold parameter θ. For instance, if the population size is sufficient for accurate estimation, then a few dependencies may help better navigate the search space even if independent variables would eventually suffice. Hence in such cases we may find a small θ value more effective even if the problem is separable. Conversely, if the model complexity (group size c) is relatively large in comparison to what can be reliably estimated from the available population of selected points, then a large θ value has the effect of forcing more variables into the set of weakly independent ones, thereby preventing overfitting. In such cases we may expect better performance with a large θ rather than a small θ value even if the problem is non-separable.

TABLE II: Ranksum statistical test for performance comparison between the use of MI and correlation coefficient in EDA-MCC for functions $F_1 - F_5$. $D = 100$, $M = 300$ and Budget size is $1 \times 10^4 \times D$

Prob.	EDA-MCC		EDA-MCC-MI		H	P-Value
	mean	std	mean	std		
F_1	1.8058	5.2604	8.4511e-004	0.0042	0	0.0653
F_2	3.7080e+004	7.4889e+004	**8.1171e+003**	**1.2053e+004**	1	0.0012
F_3	36.5014	5.9096	34.3089	6.0433	0	0.1594
F_4	0.0613	0.2281	3.1716e-004	0.0016	0	0.8380
F_5	5.7014e+006	6.6216e+006	4.8068e+006	4.8672e+006	0	0.4971

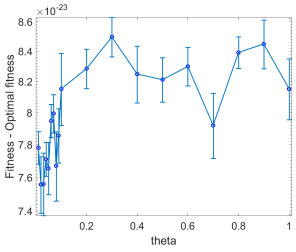
TABLE III: Ranksum statistical test for performance comparison between the use of MI and correlation coefficient in EDA-MCC for functions $F_1 - F_5$. $D = 100$, $M \in \{1000, 2000\}$ and Budget size is $1 \times 10^4 \times D$

Prob.	Population Size	EDA-MCC		EDA-MCC-MI		H	P-Value
		mean	std	mean	std		
F_1	1000	8.2679e-023	5.9495e-024	8.0915e-023	6.0637e-024	0	0.2444
	2000	**7.4054e-021**	**7.4115e-022**	7.8597e-021	8.0209e-022	1	0.0457
F_2	1000	25.0259	98.5899	6.5915	31.9113	0	0.7415
	2000	9.5780e-017	7.7764e-018	9.4713e-017	8.9666e-018	0	0.6004
F_3	1000	15.9193	2.3510	14.9642	2.3419	0	0.0814
	2000	11.2231	1.5067	11.3823	1.7490	0	0.9073
F_4	1000	9.6080e-013	2.1734e-014	9.5667e-013	2.3218e-014	0	0.7628
	2000	1.3716e-011	6.0775e-013	1.4040e-011	6.4380e-013	0	0.0667
F_5	1000	3.4448e+005	2.2463e+005	2.6968e+005	2.0549e+005	0	0.1870
	2000	1.4387e+005	1.0913e+005	1.6066e+005	1.3240e+005	0	0.7710

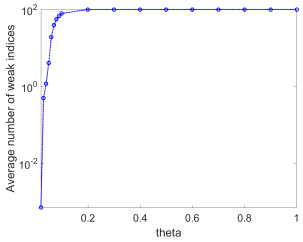
TABLE IV: Ranksum statistical test for performance comparison between the use of MI and correlation coefficient in EDA-MCC for functions $F_1 - F_5$. $D = 500$, $M = 1000$ and Budget size is $1 \times 10^4 \times D$

Prob.	EDA-MCC		EDA-MCC-MI		H	P-Value
	mean	std	mean	std		
F_1	0.0133	0.0596	7.2584e-022	1.6225e-023	0	0.5792
F_2	1.0491e+004	1.8153e+004	**1.8844e+003**	**4.1694e+003**	1	7.3522e-004
F_3	0.0030	0.0111	1.4536e-012	1.1452e-014	0	0.1091
F_4	188.4452	12.6540	**173.9188**	**8.2498**	1	1.6501e-005
F_5	6.2552e+006	3.8602e+006	**3.1108e+006**	**2.4836e+006**	1	4.2399e-005

A. *Separable Problems*



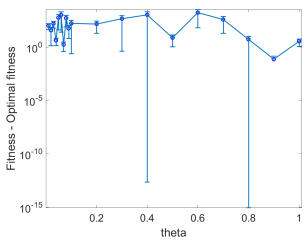
(a) Average fitness



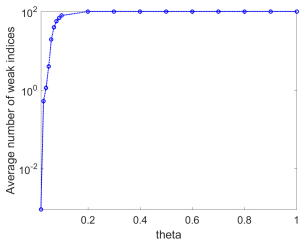
(b) Average number of weak variables

Fig. 5: Effects of $theta$ on F_1 (shifted Sphere). Average fitness achieved (*left*) and the corresponding number of variables that are treated as independent (i.e. in the set $\mathcal{W}$), as θ is varied. $D = 100$, $M = 1000$ and Budget size is $1 \times 10^4 \times D$.

Figure 5 shows the average fitness and the corresponding number of elements in $\mathcal{W}$ as θ changes, for the Shifted Sphere function. We see that EDA-MCC-MI produces its best result when $\theta = 0.03$. This turned out different from our hypothesis above. However, since shifted sphere is a quite easy problem, we observe that the fitness values are at the magnitude of 10^{-23} for all choices on θ. Therefore, for the shifted sphere function, even though $\theta = 0.03$ has the best result, other values are also good.



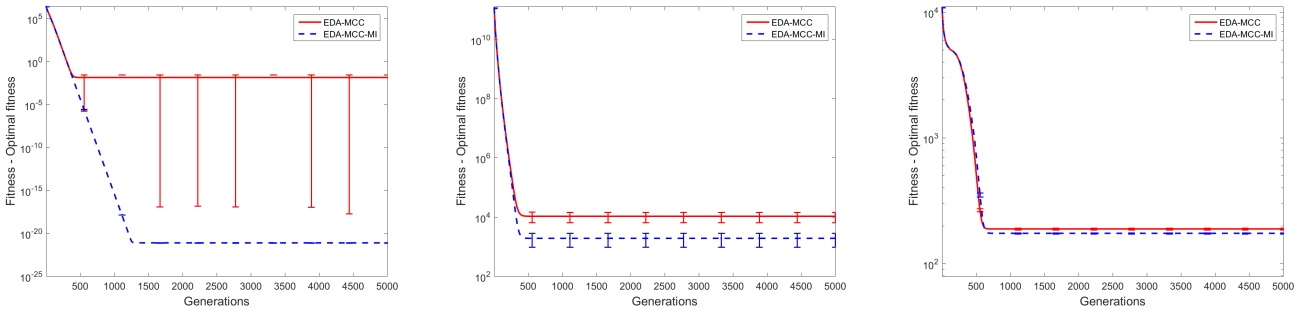
(a) Average fitness



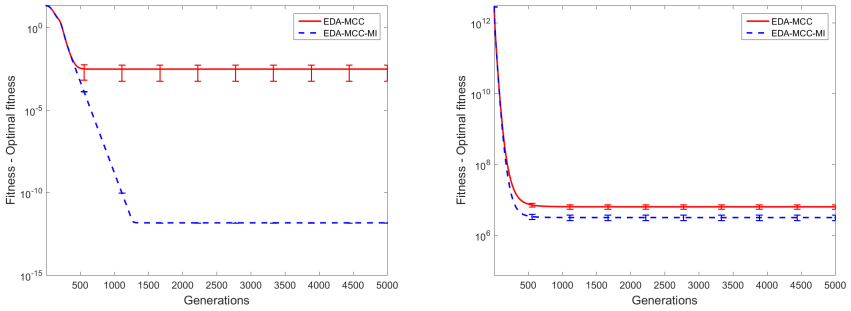
(b) Average number of weak variables

Fig. 6: Effects of $theta$ on F_2 (shifted Elliptic). Average fitness achieved (*left*) and the corresponding number of variables that are treated as independent (i.e. in the set $\mathcal{W}$), as θ is varied. $D = 100$, $M = 1000$ and Budget size is $1 \times 10^4 \times D$.

Likewise, Figure 6 shows the average fitness and the corresponding number of elements of $\mathcal{W}$, as θ changes, for the Shifted Elliptic function. When $\theta \geq 0.2$, all the variables are in $\mathcal{W}$ (i.e. weakly correlated), therefore the results with any larger θ value should be almost the same. In turn, the results slightly fluctuate, and the error bars are wide, which means



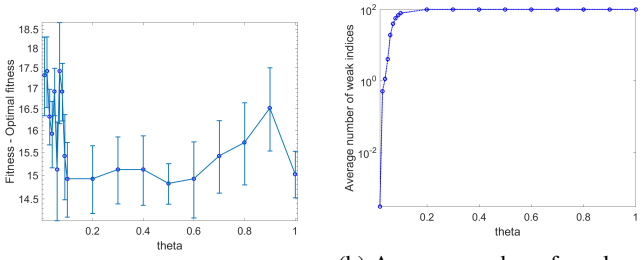
(a) F_1 : Shifted Sphere (b) F_1 : Shifted Sphere (c) F_1 : Shifted Sphere



(d) F_3 : Shifted Ackley (e) F_1 : Shifted Sphere

Fig. 4: Evolutionary curves for performance comparison between the use of MI and correlation coefficient in EDA-MCC for functions F_1 and F_4. $D = 500$, $M = 1000$ and Budget size is $1 \times 10^4 \times D$

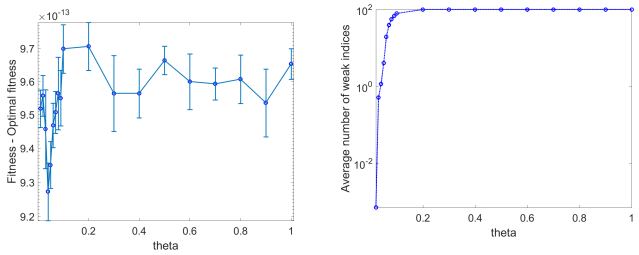
that, there are no statistically significant differences here. All values of θ perform similarly.



(a) Average fitness (b) Average number of weak variables

Fig. 7: Effects of $theta$ on F_3 (Rastrigin). Average fitness achieved (*left*) and the corresponding number of variables that are treated as independent (i.e. in the set $\mathcal{W}$), as θ is varied. $D = 100$, $M = 1000$ and Budget size is $1 \times 10^4 \times D$.

The results of analogous experiments on the Rastrigin function are given in Figure 7. This result agrees with our hypothesis. We see that when $\theta > 0.1$, EDA-MCC-MI generally achieves good results. The Rastrigin function is indeed separable and the best performing range of θ values does indeed place all variables in the weakly dependent category. However, the slight fluctuations and large error bars when $\theta > 0.1$, suggest that more experiments should be done on this problem, in order to obtain statistically more conclusive results.



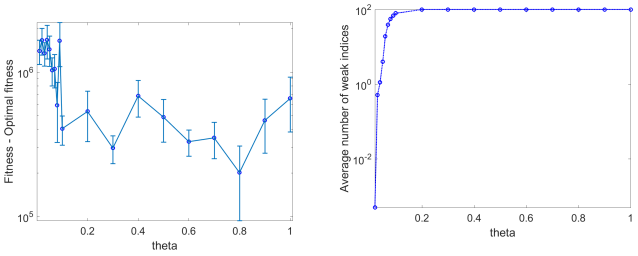
(a) Average fitness (b) Average number of weak variables

Fig. 8: Effects of $theta$ on F_4 (Ackley). Average fitness achieved (*left*) and the corresponding number of variables that are treated as independent (i.e. in the set $\mathcal{W}$), as θ is varied. $D = 100$, $M = 1000$ and Budget size is $1 \times 10^4 \times D$.

Figure 8 shows the average fitness and the corresponding number of elements in $\mathcal{W}$, as θ changes, for the Ackley function. We can see that EDA-MCC-MI obtains its best result when $\theta = 0.04$. This is a little surprising, as very few weak variables are used to estimate univariate models and many strongly dependent variables are used to generate multivariate models. It may be that modelling of dependencies helps in the intermediate stages of the search here. However, we also notice that the Ackley function is in fact somewhat similar to Sphere, in terms of its overall shape, apart from having local optima. Indeed, all the fitness values are at the magnitude of 10^{-13}, which is very close to the global optimum of zero.

Therefore the choice of θ does not make a big difference on this function after all.

B. Non-Separable Problem



(a) Average fitness

(b) Average number of weak variables

Fig. 9: Effects of $theta$ on F_5 (Rosenbrock). Average fitness achieved (*left*) and the corresponding number of variables that are treated as independent (i.e. in the set $\mathcal{W}$), as θ is varied. $D = 100$, $M = 1000$ and Budget size is $1 \times 10^4 \times D$.

Finally, Figure 9 depicts the average fitness and the corresponding number of elements in $\mathcal{W}$ as θ changes on Rosenbrock function. We see that when $\theta \geq 0.2$, EDA-MCC-MI has better results than in the case of smaller values of θ. This disagrees with our initial hypothesis, because we expected that θ should be small for this problem in order to keep more dependencies. Yet, in this case, the results turns out better when more variables are used to estimate univariate models and fewer variables are used to generate multivariate models. This is somewhat surprising, and warrants further investigations in future work. It is possible that the multivariate models overfit, or it may be – noticing that the best fitness found is relatively far from the global optimum – that the suboptimal result reached is easily reachable with independent variables. We should note however, that the differences seen in Figure 9 were not found to be statistically significant.

C. Further remarks

Taking a look at all figures presented in this section, it may be worth observing that the rightmost plots in all of these figures look almost the same. This means that θ causes similar splitting proportions in all the functions tested.

Therefore, based on the results in Figures 5-9 together, our exploratory experiments suggest that in the case of Sphere and Ackley our method works best when θ is 0.03 or 0.04, when only a few variables are classified as weak variables, and in the case of the other three functions we can achieve better results when θ is greater than 0.1, which means that more independent variables are beneficial. However, the large error bars suggest that more experiments would be needed for a definite conclusion.

VI. Conclusion

We proposed and investigated an extension to a recent method for large scale EDA search. The method controls the complexity of the search distribution by estimating the essential dependency structure of the search variables and discarding weak dependencies. Our proposal is to replace linear correlation coefficient estimates by mutual information (MI) estimates. Our method is in a better position to determine the correct dependency structure than the EDA-MCC can do. Although in our experiments we only used a Gaussian search distribution, which has a very limited ability to make use of the improved dependence structure estimates, our empirical comparison results demonstrated that our approach, EDA-MCC-MI is never worse than EDA-MCC and performs better in most cases. Our implementation employs a nonparametric MI estimator, hence the use of MI can easily be combined with more powerful non-Gaussian search distributions in further research.

Acknowledgment

This work was performed as part of the first author's MSc mini-project at the University of Birmingham.

References

[1] P. Larranaga and J. A. Lozano, *Estimation of distribution algorithms: A new tool for evolutionary computation*, vol. 2. Springer Science & Business Media, 2002.

[2] E. Zitzler, K. Deb, and L. Thiele, "Comparison of multiobjective evolutionary algorithms: Empirical results," *Evolutionary computation*, vol. 8, no. 2, pp. 173–195, 2000.

[3] M. Sebag and A. Ducoulombier, "Extending population-based incremental learning to continuous search spaces," in *Proc. of Parallel Problem Solving from Nature (PPSN)*, pp. 418 – 427, Springer, 1998.

[4] M. Wagner, A. Auger, and M. Schoenauer, "EEDA: A new robust estimation of distribution algorithm," *Research Report RR-5190 INRIA*, 2004.

[5] M. L. Sanyang and A. Kabán, "Multivariate Cauchy eda optimisation," in *Intelligent Data Engineering and Automated Learning–IDEAL 2014*, pp. 449–456, Springer, 2014.

[6] J. H. Friedman, *An overview of predictive learning and function approximation*. Springer, 1994.

[7] A. Kabán, J. Bootkrajang, and R. J. Durrant, "Towards large scale continuous eda: a random matrix theory perspective," *Evolutionary computation*, 2015.

[8] M. L. Sanyang and A. Kabán, "Heavy tails with parameter adaptation in random projection based continuous eda," in *Proc. of the IEEE Congress on Evolutionary Computation (CEC)*, pp. 2074 –2081, 2015.

[9] L. de la Ossa, J. Gamez, and J. Puerta, "Migration of probability models instead of individuals: an alternative when applying the island models to edas," in *Proc. of Parallel Problem Solving from Nature (PPSN), LNCS vol. 3242*, pp. 242–252, 2004.

[10] W. Dong, T. Chen, P. Tino, and X. Yao, "Scaling up estimation of distribution algorithms for continuous optimization," *Evolutionary Computation, IEEE Transactions on*, vol. 17, no. 6, pp. 797–822, 2013.

[11] P. Larranaga, R. Etxeberria, J. A. Lozano, and J. M. Pena, "Combinatorial optimization by learning and simulation of bayesian networks," in *Proc. 16th Annu. Conf. Uncertainty Artif. Intell. (UAI-2000)*, pp. 343–352, 2000.

[12] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[13] G. A. Darbellay, I. Vajda, *et al.*, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[14] T. Marek, P. Tichavsky, G. Dohnal, *et al.*, "On the estimation of mutual information," *J. A, Dohnal G, editors*, 2008.

[15] P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y.-P. Chen, A. Auger, and S. Tiwari, "Problem definitions and evaluation criteria for the cec 2005 special session on real-parameter optimization," *KanGAL report nr. 2005005*, 2005.

[16] M. L. Sanyang, R. J. Durrant, and A. Kabán, "How effective is Cauchy eda in high dimensions?," in *Proc. of the IEEE Congress on Evolutionary Computation (IEEE CEC 2016)*.