An Estimation Distribution Algorithm with the Spearman's Rank Correlation Index

Arturo Hernández Aguirre, Enrique Villa Diharce, Selma Barba Moreno
Centre for Research in Mathematics, Department of Computer Science
AP. 402, Guanajuato, Gto.
CP. 36000, México
artha, villadi, selma@cimat.mx

ABSTRACT

This article arguments that rank correlation coefficients are powerful association measures and how can they be adopted by EDAs. A new EDA implements the proposed ideas: the Non-Parametric Real-valued Estimation Distribution Algorithm (NOPREDA). The paper fully describes the rank correlation coefficient, and the procedure to build a non parametric model for the probability distribution of the source data. A benchmark of global optimization problems is solved with NOPREDA.

Categories and Subject Descriptors: J.2[Physical Sciences and Engineering]Mathematics and Statistics

General Terms: Algorithms, Design, Performance.

Keywords: Estimation Distribution Algorithms, Rank Correlation.

1. INTRODUCTION

Correlation is one measure of the strength of association or dependence between two variables. In the parametric model context, the Pearson product moment correlation coefficient estimates the degree of *linear* association between two variables (this is the common correlation index found in most introductory books). The correlation coefficient of two variables X and Y is computed through the variances σ_{xx} and σ_{yy} and covariance σ_{xy}, as follows:

$$r = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}} \qquad (1)$$

The coefficient r takes values between -1 and $+1$. If large values of one variable occur with small values of the other the correlation is negative. If the two variables are independent, the value of the coefficient is zero. However, a nonlinear relationship can also produce a correlation value close to zero.

Other measures of bivariate association are for example, the Spearman rank correlation coefficient, represented by ρ, and the Kendall rank correlation coefficient represented by τ. These coefficients are confined to the interval $(-1, 1)$, with the following meaning at critical values: a lack of association is represented by a value near zero. Values near $+1$ imply a strong positive association and values near -1 imply a strong negative association. However, for the Pearson coefficient, $r = \pm 1$ implies linearity but for the rank coefficients a value of ± 1 usually do not imply linearity. Rather, they imply monotonicity. If two variables increase together, this is a monotonic increasing relationship, whereas if one increases and the other decreases, the relationship is monotonic decreasing. Exact value of $+1$ or -1 imply the relationship is strictly increasing monotonicity or strictly decreasing monotonicity. For a simple example of monotonicity note that the Spearman's ρ of the pairs (X, Y), such that $X \sim N(0, 1)$, and $Y = X^3$ is $\rho(X, Y) = +1$. However, the Pearson correlation coefficient is around 0.8 Another example, $X \sim N(0, 1)$ and $Y = 2 * X$, both Spearman and Pearson correlation coefficients are $+1$). Note that in both examples the Spearmant's ρ coefficient correctly indicates the association due to *Y value increases if X value increases.*

The goal of this paper is twofold. First and most important, to present the goodness of a rank correlation coefficient, and a way to use this association measure in a practical implementation called the Non-parametric Real-valued Estimation Distribution Algorithm (NOPREDA). Second, to provide an empirical assessment of its capacity to solve global optimization problems.

2. THE NOPREDA ALGORITHM AND EXPERIMENTS

A pseudo-code of NOPREDA is shown in Figure 1. NOPREDA creates a model of the PDF in two steps: first, it computes the cumulative empirical distribution (CED) of each variable. Second, it computes the Spearman rank correlation matrix of the sample set. Simulating new data from this model (to populate the next generation) also requires two steps: first, new data is simulated from every CED; second, the rank correlation matrix is induced in the new data. To achieve this step, we do use a procedure introduced by Iman and Conover to simulate new data regardless of the distribution with a specified Spearman's rank correlation index [2].

2.1 Experiment 1

For this experiment the goal is to reach the optimum within 1E-6. However, there is a limited effort to spend in

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

```
NOPREDA
t=0;
$P_t$ ← InitialPopulation;
Repeat
  $P_t$ ← evaluatefitness($P_t$);
  S ← select_best($P_t$);
  IDX ← find_indexes_toinducecorrelation(S);
    For each variable $X_i \in S$
      $CD_i$ ← compute_cumulative_distribution($X_i$);
      New$X_i$ ← generate_newpopulation($CD_i$);
    EndFor
  t=t+1;
  $P_t$ ← {New$X_i$}; % $nextpopulationisamatrix$
  $P_t$ ← reorder($P_t$, IDX);
Until Termination;
function find_indexes_toinduce_rank correlation(S)
  %This is Iman and Conover Algorithm
  RCM ← compute_rankcorrelation_matrix(S);
  CHO ← Choleskyfactorization(RCM);
  T1 ← Any uncorrelated data set ;
  T2 ← T1 × $CHO^T$ ;
  IDX ← find_sorting_index(T2);
return IDX;
```

Figure 1: Main pseudocode of NOPREDA

the process, hence the number of function evaluations is 3E5. A run is stop if stagnation is detected. An improvement smaller than 1E-7 in the last 10,000 function evaluations means stagnation. The optimum value of all these problems is 0.0 The population size is 30, truncation selection picks the best 20 elements, the minimum allowed variance factor is 1E-5. Elitism of 2 individuals. Results for this experiment are shown in Table 1.

Problem	Best Approximation	Evaluations
Dimension 10		
Sphere	8.5E-07 ± 1.6E-07	8874 ± 875
Rosenbrock	9.1E-01 ± 9.5E-01	3E05 ± 0.0
Griewank	8.0E-07 ± 1.5E-07	9118 ± 825
Ackley	9.1E-07 ± 1.0E-07	13873 ± 1296
Rastrigin	8.1E-07 ± 1.9E-07	16822 ± 3224
Dimension 50		
Sphere	9.4E-07 ± 4.4E-08	81774 ± 1736
Rosenbrock	82.1 ± 32.5	3E05 ± 0.0
Griewank	9.5E-07 ± 3.2E-08	82353 ± 2077
Ackley	9.7E-07 ± 2.7E-08	214995 ± 13882
Rastrigin	6.3E-07 ± 2.6E-07	249600 ± 23950

Table 1: NOPREDA results for Experiment 1

2.2 Experiment 2

The functions of this experiment are convex and monotone, reported by [1]. The aim of the experiment is to observe the scalability of the algorithm. The average number of evaluations is measured while the dimensionality of the problem takes the values 2,4,8,10,20,40 and 80. The functions are defined in Table 2.

Figure 2 shows the results for NOPREDA. The largest dimension in which the problems were solved is 40. This is basically due to lengthly computation required for dimension 80.

Name	Definition	Value to reach
Sphere	$\sum_{i=1}^{l} x_i^2$	10^{-10}
Ellipsoid	$\sum_{i=1}^{l} 10^{6\frac{i-1}{l-1}} x_i^2$	10^{-10}
Cigar	$x_1^2 + \sum_{i=2}^{l} 10^6 x_i^2$	10^{-10}
Tablet	$10^6 x_1^2 + \sum_{i=2}^{l} x_i^2$	10^{-10}
Cigar Tablet	$x_1^2 + \sum_{i=2}^{l-1} 10^4 x_i^2 + 10^8 x_l^2$	10^{-10}
Two Axes	$\sum_{i=1}^{\lfloor l/2 \rfloor} 10^6 x_i^2 + \sum_{i=\lfloor l/2 \rfloor}^{l} x_i^2$	10^{-10}
Different Powers	$\sum_{i=1}^{l} \lvert x_i^2 \rvert^{2+10\frac{i-1}{l-1}}$	10^{-15}

Table 2: Test functions and values to reach of Experiment 2

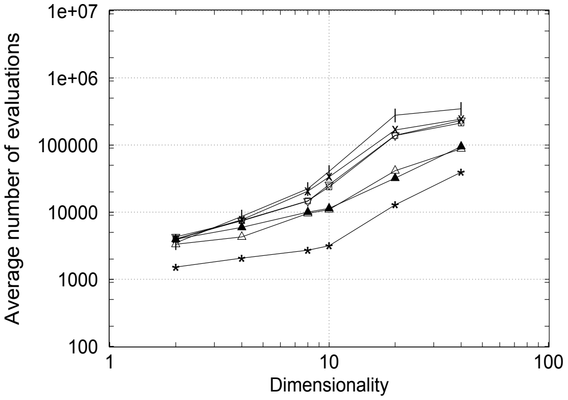

Figure 2: NOPREDA plots for Average Number of Evaluations vs Dimensionality (Functions solved up to dimension 40) Cigar + , Cigar tablet ×, Different powers *, Ellipsoid □, Sphere △, Tablet ▲, Two axes ▽.

3. CONCLUSIONS

Non parametric approaches should be studied in the context of EDAs because of their ability to create good approximations of the PDF. Further, the Spearman's correlation coefficient is presented as a better indicator of a non linear relation.

Acknowledgements

The authors gratefully acknowledge support from CONACyT project No. 51667-Y.

4. REFERENCES

[1] J. Grahl, P. A. N. Bosman, and F. Rothlauf. The correlation-triggered adaptive variance scaling IDEA. In *GECCO '06: Proceedings of the 8th annual conference on Genetic and Evolutionary Computation*, pages 397–404. ACM Press, 2006.

[2] R. L. Iman and W. J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics: Simulation and Computation*, 11(3):311–334, 1982.