(1+2)-Evolution Strategy for Fitting a Straight Shuffle of Min to a Dataset *

Alfredo Cuesta-Infante
School of Computer Science
Felipe II College
Aranjuez, Spain
acuestai@pdi.ucm.es

José Ignacio Hidalgo
School of Computer Science
Univ. Complutense de Madrid
Spain
hidalgo@dacya.ucm.es

María Victoria Rivas
School of Business
Felipe II College
Aranjuez, Spain
mariavictoria.rivas@pdi.ucm.es

ABSTRACT

This paper proposes an Evolution Strategy to find a shuffle of M that suits to a batch of datasets that serve as benchmark. Results are compared with wavelet estimation.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Evolution Strategies and Evolutionary Programming*

General Terms

Algorithms

Keywords

Evolution strategies, Copula functions, Shuffles of M, Probability distribution functions, Non-parametric estimation.

1. INTRODUCTION

Due to Sklar's theorem any bivariate joint cumulative probability distribution function (JCDF) $F_{XY}(x, y)$ can be decomposed in a copula function C that describes the scale-free dependence structure and two margins $F_X(x)$ and $F_Y(y)$ that model the distribution of each random variable involved [9]. In other words

$$F_{XY}(x, y) = C(u, v) \quad , \quad u = F_X(x) \quad , \quad v = F_Y(y).$$

This separation is the reason why copula functions are gaining an increasing interest in diverse fields. A recent compilation of applications can be found in [3]. A comprehensive introduction to copula functions can be found in [4] and [8].

Any copula can be approximated by a *Shuffle of M* , being M the copula known as *Fréchet-Hoeffding upper bound* $M(u, v) = \min(u, v)$ [6]. Shuffles of M are simply characterized by a permutation of the array $[0, \ldots, n-1]$.

The importance of shuffles of M has been highlighted in [1, 5, 7] for instance. Motivated by its potential interest, this paper aims to find the permutation such that the resulting shuffle of M fits to a given dataset $\{(u_i, v_i)\}_{i=1,2,\ldots,N_s} \in [0, 1]^2$. Since n different elements yield to $n!$ possible permutations this work presents an evolution strategy to carry out the search. Outcomes are compared with wavelet based estimation of the empirical copula in order to validate the proposal.

*This work has been partially supported by Spanish Government grants TIN2008-00508, MEC Consolider Ingenio CSD00C-07-20811 of the Spanish Council of Science and Technology and Programa José Castillejo JC2008-00421.

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

2. EVOLUTION STRATEGY PROPOSAL

In our proposal a single individual gives birth to two descendants and, after a tournament scoring their fitness only the best one remains. It is therefore a (1+2)-ES.

Representation. Each shuffle of M is to be considered an individual of the solution space. A *straight* shuffle of M (the only one considered here) is characterized by a permutation without repetition of the array $[0, 1, \ldots, n-1]$. We will call *chromosome* to the array and *gene* to one element in it.

Split in two. A chromosome is splited in two ways, each one carrying half of the genes and erasing the other half.

Recombination. Each offspring completes the empty half with the genes of the another half in a random order. Hence offspring's representation will always be correct.

Mutation. Mutation implies replacement of one gene from the transmited half with one gene of the completed half. Genes to be exchanged are chosen randomly and the mutation happens with probability p_m, which is a parameter of the algorithm.

Fitness. The starting point is a 3D histogram giving the relative frequency of the pairs $\{(u_i, v_i)\}$, with $i = 1, 2, \ldots, N_s$. Usually the number of bins is $n \times n$, where $n = 2^J$ and $J = \lfloor \log_2 \sqrt{N_s} \rfloor$. By means of the cumulative sum of the relative frequencies in the histogram, the empirical copula $\overline{C}$ is obtained. Let $\widehat{C}$ be the individual being tested, then the estimation of the fitness that we propose is:

$$f = \sum_{j,k=1}^{n} \epsilon_{j,k}^2, \text{ with } \epsilon_{j,k} = \widehat{C}(s_j, t_k) - \overline{C}(s_j, t_k), \quad \text{for } s_j = \tfrac{j-1}{n-1} \text{ and } t_k = \tfrac{k-1}{n-1}. \tag{1}$$

No extra constraints are needed.

Stop condition. The algorithm ends either when a minimum value of the fitness (f_{ϵ^*}) has been attained, or when a maximum number of generations max_T have been completed. The former is provided assuming that the difference

$$\widehat{C}(s_j, t_k) - \overline{C}(s_j, t_k) = \epsilon^*$$

for every pair (s_j, t_k) evaluated. Thus, ϵ^* is a fixed expected value of the difference between the estimated and the em-

Table 1: Results of the batch

Code	a	μ_1	μ_2 $\times 10^{-2}$	μ_3 $\times 10^{-4}$	μ_4 $\times 10^{-4}$
C1.	0.5	0.129	2.694	2.181	8.306
C2.	2	0.123	2.1	2.288	12.3
C3.	10	0.212	10.98	2.734	2.534
G1.	1	0.255	15.31	2.999	2.912
G2.	2	0.296	19.37	3.323	3.664
G3.	5	0.099	−0.276	2.263	8.531
F1.	−12	0.121	1.82	2.3	16.19
F2.	−6	0.171	6.813	2.512	15.24
F3.	−2	0.235	13.27	2.794	7.385
F4.	2	0.202	9.996	2.701	2
F5.	6	0.177	7.483	2.473	2.306
F6.	12	0.173	7.011	2.536	2.337
X1.	−1	0.244	24.39	3.94	2.94
X2.	1	0.258	25.83	4	3

C=Clayton, G=Gumbel, F=Frank, X=Eq.(3)
a=parameter of the copula.

pirical copula; which leads to

$$f_{\epsilon^*} = \epsilon^{*2} n^2. \quad (2)$$

3. PERFORMANCE AND DISCUSSION

For the sake of keeping the focus of this paper on the algorithm, hereafter only pairs $\{(u_i, v_i)\}_{i=1,\ldots,N_s}$, both u_i and v_i uniformly distributed, are considered. Thus, the problem of estimating the margins is left aside as an statistical issue.

We present a batch of 12 datasets of $N_s = 1024$ iid pairs drawn from three different copulas (Clayton, Gumbel and Frank) plus 2 datasets drawn from the copula

$$C(u,v) = uv + a(2u-1)(2v-1)(u-1)(v-1)uv, \quad (3)$$

All together they represent a wide range of dependence structures and are listed in Table 1. The choice of $N_s = 2^{10}$ leads to $n = 32$. The value of ϵ^* for all the cases was set to 10^{-2} so $\epsilon^{*2} = 10^{-4}$ and $f_{\epsilon^*} = 0.1024$. Such a demanding value was selected in order to complete 1000 generations so a study of the convergence can also be made. In addition a non-parametric estimation of the copula following the recent procedure given in [2] is done in order to compare and validate the method here proposed.

According to the good practice in heuristic optimisation, the algorithm was ran ten times for every copula of the batch. Each iteration i produces the following elements: f_i is the value of the best fitness at the end of the execution i, $\Delta f_i = f_i - f_{\epsilon^*}$ is the deviation of the prior from the maximum value imposed, $h_i(\epsilon^2_{j,k})$ is the distribution of $\epsilon^2_{j,k}$ obtained from a histrogram of ten bins within the interval $[0,\ 3{\cdot}10^{-3}]$ (the cumulative distribution is then $H_i(\epsilon^2_{j,k})$), $S_i = E_{h_i}(\epsilon_{j,k})^2$ is the expected value according to the distrubution h_i of the squared differences between the shuffle of M obtained (the estimated copula) and the empirical copula, and finally $\Delta S_i = S_i - \epsilon^{*2}$ is the deviation of the prior from the maximum value imposed.

Then the following averages are computed for each copula:

$$\mu_1 = \mu(\{f_i\}), \quad \mu_2 = \mu(\{\Delta f_i\}), \quad \mu_3 = \mu(\{\Delta S_i\}), \quad (4)$$

for $i = 1, \ldots, 10$, and where the operation $\mu(\mathbf{x})$ returns the mean of $\mathbf{x}$. Results are shown in columns μ_1, μ_2 and μ_3, of Table 1. Variances are all at least two orders of magnitude smaller than its corresponding mean; thus, for the sake of clariry they do not appear in the table.

In addition a non-parametric wavelet estimation using Daubechies-8 second approximation as proposed in [2] has been done. The estimated copula obtained with this method will be denoted $\widehat{C}_W$. Consider the distribution of the squared differences bewteen $\widehat{C}_W$ and $\overline{C}$, denoted by

$$h_W\left(\left(\widehat{C}_W(s_j,t_k) - \overline{C}(s_j,t_k)\right)^2\right).$$

Then constructing the histogram of ten bins within the same interval than $h(\epsilon^2_{j,k})$ it is possible to compute the expected value

$$\mu_4 = E_{h_W}\left(\left(\widehat{C}_W(s_j,t_k) - \overline{C}(s_j,t_k)\right)^2\right). \quad (5)$$

Results of (5) are shown in column μ_4, in Table 1.

Comparison and Validation. From column μ_2 in Table 1 a first classification can be made, considering those cases with $\mu_2 \geq 0.1$ and the rest of them, i.e. a different order of magnitude in the expected deviation from f_{ϵ^*}. The latter shows a very good outcome of the fitness at the end of the execution. Moreover, despite the demanding value of f_{ϵ^*}, it has been attained by G3, with $\mu_1 = 0.099$. Concerning the former it is interesting to separate C1, G1, F3, F4 from G2. The first group are those close to the limit case in which Clayton, Gumbel and Frank copulas turn into $\Pi(u,v) = uv$ (independence). Thus, it is clear that they will all give similar results. In addition, performance in X1 y X2 are similar to the first group. Although different from $\Pi(u,v)$, their density is quite spread in the unit square; hence results are close to the worst.

Comparing columns μ_3 and μ_4, wavelet estimation performs clearly better only for F4. For the rest of cases close to copula Π it performs similarly, such as C3 or G1, or even much worse than the proposal as in F1.

4. REFERENCES

[1] A. Erdely, J. Gonzalez-Barrio, and R. Nelsen. Symmetries of random discrete copulas. *Kybernetika*, 44(6):846–863, 2008.

[2] C. Genest, E. Masiello, and K. Tribouley. Estimating copula densities through wavelets. *Insurance: Mathematics and Economics*, 44(2):170–181, 2009.

[3] P. Jaworski, F. Durante, W. Härdle, and T. Rychlik (editors) *Workshop on Copula Theory and its Applications, Lecture Notes in Statistics - Proceedings.* Springer-Berlin, 2010.

[4] H. Joe. *Multivariate models and dependence concepts.* Chapman and Hall, 1997.

[5] G. Mayor, J. Suñer, and J. Torrens. Copula-like operations on finite settings. *IEEE Transactions on Fuzzy Systems*, 13(4):468–477, 2005.

[6] P. Mikusinski, H. Sherwood, and M. Taylor. Shuffles of min. *Stochastica*, XIII-1:61–74, 1992.

[7] P. Mikusinski and M. Taylor. Some approximations of n-copulas. *Metrika*, 2009.

[8] R. B. Nelsen. *An introduction to copulas.* Springer Series in Statistics, 2nd. edition, 2006. 2

[9] A. Sklar. Fonctions de repartition n dimensions et leurs marges. volume 8. Inst. Statist. Univ. Paris, 1959.