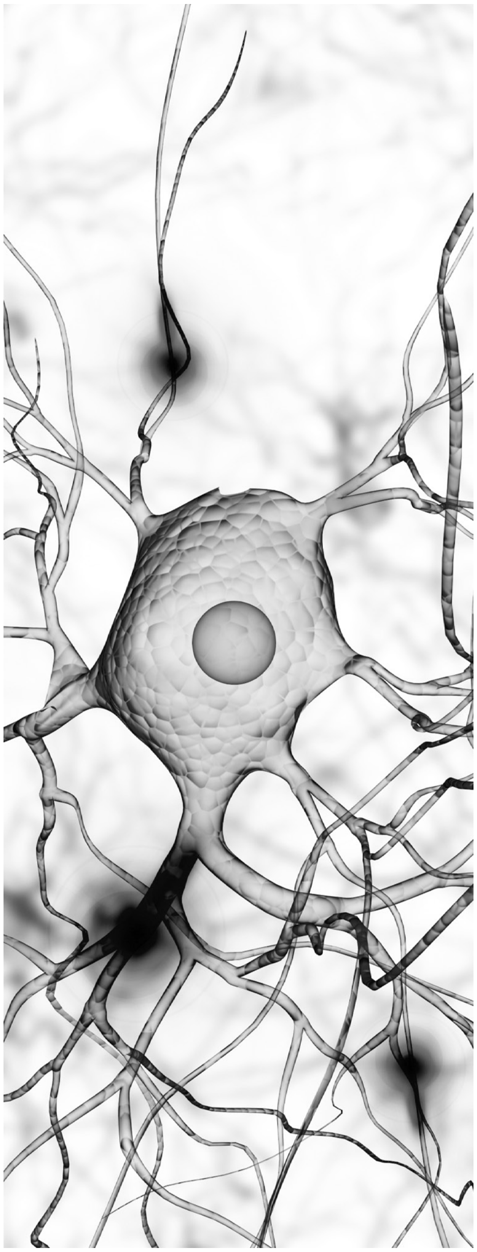
Using genetic algorithms as a parameter estimation tool for the generalized lambda distribution (gld) family: “methods of moments”

David Leonardo Moreno Bedoya y Nelson Ricardo Fino Puerto*



Abstract

The generalized lambda distribution, $GLD\ (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is a four-parameter family that has been used for fitting distributions to a wide variety of data sets. Minimization through traditional calculus-based methods has been implemented with relative success, but due to computational and theoretical shortcomings of those methods, the moment space has been limited. This paper solve those troubles by using Genetic Algorithms (search algorithms based on the mechanics of natural selection and natural genetics) applied to the methods of moments. Examples of better solutions than the ones find out with traditional calculus-based methods are included.

Key words and phrases: Data Fitting; Generalized Lambda Distribution; Minimization Method; Moments, Percentiles, Genetic Algorithms.

Introduction

In sciences such as decision analysis, it is desirable to represent data from real situations through a known distribution. Many researchers around the subject have chosen the generalized lambda distribution(GLD)[1] as a distribution that it is able to provide a flexible approach to data fitting. Traditionally, finding GLD parameters through the method of moments involves great computational efforts providing low accuracy in the results. One of the goals of optimization, is improvement[2]. With traditional calculus-based methods it is possible to reach a single unique solution in only one direction. Now, Imagine an algorithm that is computationally simpler and equally or maybe more powerful in its search for improvement, a method that it is able to jump between all possible hills or valleys in order to find the best possible solution. This algorithm is named Genetic Algorithms and they are going to take a principal role in this investigation along with GLD.

In section 2, we provide a background of the GLD family and the method of moments to show the dimension of the problem, in section 3, we talk about GA their advantages against traditional calculus-based methods and the reasons why we have chosen this methodology, section 4 will provide an example

Recibido: 30 octubre de 2010
Aceptado:12 mayo de 2011

* Docentes Facultad Ing Sistemas. Universidad Antonio Nariño. Correo electrónico respectivamente Nrfino@eudoramail.com y Dmoren0@hotmail.com

that will show and support our hypothesis and finally in section 5 we provide some conclusions.

GLD family

The GLD family of distributions is defined through its inverse distribution function [1] by

$$Q\,(y) = Q\,(y;\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (\,1-y)^{\lambda_4}}{\lambda_2} \quad (1)$$

Where $y \;\; u(0,1)$

This representation of *Q(y)* is particularly convenient for simulation studies and data analysis where one may wants to generate random samples from a specific $GLD\,(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ distribution. The $GLD\,(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is valid if and only if (λ_3, λ_4) is in one of the regions marked *1,2,3,4,5* or *6* in figure 1; provided that λ_2 has the same sign as

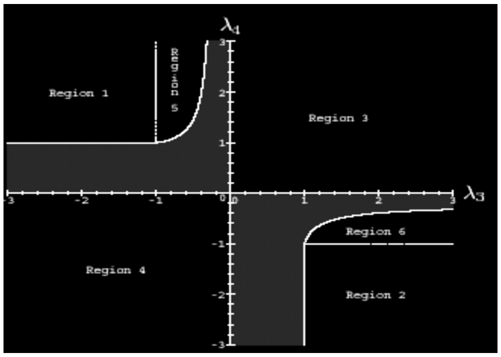
$$g\,(y, \lambda_3, \lambda_4) \equiv \lambda_3\; y^{\lambda_3}-1 + \lambda_4\;(1-y)^{\lambda_4}-1$$

for all ***y*** in the interval [0,1].

A way of fitting a specific $GLD\,(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ to $X_1, X_2, \ldots, X_{n-1}, X_n$ a given data set, is through the method of moments where the sample moments are defined [6] by

$$\alpha_1 \;\; = \mu = E(X) = \lambda_1 + \frac{A}{\lambda_2},$$
$$\alpha_2 = \sigma^2 = E\left[(X-\mu)^2\right] = \frac{B-A^2}{\lambda_2^2}, \quad (2)$$
$$\alpha_3 = E(X-E(X))^3/\sigma^3 = \frac{(\,C-3AB+2\,A^3\,)}{\lambda_2^3\sigma^3},$$
$$\alpha_4 = E(X-E(X))^4/\sigma^4 = \frac{(\,D-4\,AC+6\,A^2B-3\,A^4\,).}{\lambda_2^4\sigma^4}$$

Figure 1. Regions 1, 2, 3,4, 5 and 6 are valid regions. The shaded region is not valid [8].



Where

$$A = 1/(1+\lambda_3) - 1/(1+\lambda_4),$$
$$B = 1/(1+2\lambda_3) + 1/(1+2\lambda_4) - 2\beta(1+\lambda_3, 1+\lambda_4), (3)$$
$$C = 1/(1+3\lambda_3) - 3\beta(1+2\lambda_3, 1+\lambda_4) + 3\beta(1+\lambda_3, 1+2\lambda_4) - 1/(1+3\lambda_4),$$
$$D = 1/(1+4\lambda_3) - 4\beta(1+3\lambda_3, 1+\lambda_4) + 6\beta(1+2\lambda_3, 1+2\lambda_4) - 4\beta(1+\lambda_3, 1+3\lambda_4)$$
$$+1/(1+4\lambda_4)$$

The $\beta(u,v)$ expression is the beta function defined as in [6] and [3]. All the $GLD\,(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ parameters found in this article, used the method of moments implemented in Algorithms 2009[7], and minimized the following optimization function:

$$f\,(\lambda_3, \lambda_4) = |\,\alpha_3 - |\,\hat{\alpha}_3\,|\,|^2 + |\,\alpha_4 - \hat{\alpha}_4|^2 (4)$$

The minimization, performed using genetic algorithms(GA), had as purpose the finding of optimal parameters for, (λ_3, λ_4), in Eq. (2), which will give us the value of A,B,C and D, and they will allow us to find the remaining parameters (λ_1, λ_2). In order to minimize Eq. (4), we find a correspondence to the data through the classical definition of moments for collected data [6]:

Note: A more detailed description of the method of moments it's available at Zaven A. Karian and Edward J. Dudewicz [6].

The stopping criteria, E, for the method of moments is defined by $E = max(|\alpha_3 - |\hat{\alpha}_3||, |\,\alpha_4 - \hat{\alpha}_4|)\,(10)$. The GA will run until, E, is found or the number of generations (cycles) is satisfied.

Genetic algorithms (GA)

Genetic algorithms are search algorithms based on the mechanics of natural selection and natural genetics. They combine survival of the fittest, among string structures, with a structured, yet randomized information exchange between individuals. The methodology forms an algorithm with some of the innovate flair of human search. As with humans, in every generation, a new set of creatures(offspring) is created using DNA(bits-strings) and pieces of information of the parents; an occasional mutation might take place such that the solution might find a different path to another minimum, hopefully, the global minimum. GA are not a simple random walk procedure. They efficiently exploit historical information to speculate on new search points with expected improved performance [2].

Genetic algorithm can be viewed as follow [4]: “They begin with a random population, then, using an objective function(which the user provides), it selects the best individuals and mate them using different kinds of methods, afterwards a new generation is created. The GA mutates the offspring and inserts it into the population. A stopping criteria is then evaluated (it can be number of generations or the error E), if the criteria is not satisfied it will return to select individuals for mating. Then, the cycle begins all over”. The GA used in the present paper are based on the GALib library, a library from MIT (Massachusetts Institute of Technology) [4]. The GA selected for the minimization task was the one described by David Goldberg (simple GA) in [2], the simple GA involves Roulette Wheel Selector and a typical genome of zeros and ones to describe an individual inside the population, it also used the classical most known operators: reproduction, crossover and mutation.

Note: For a complete reference of GALib classes you can refer to: “GALib: A C++ Library of Genetic Algorithm component”, Matthew Wall [4].

Numerical examples and results

From Ramberg, Dudewicz, Takidamalla and Mykytka(1979)[1], we can see that there are four possible regions where the solutions can be found, two of them are not available regions and one of them is an impossible region. A restriction is found on

$$1.8 + 1.7\alpha_3^2 \leq \alpha_4 \leq 1.7\alpha_3^2 + 9$$

because of table space and difficulties associated with computations when the upper restriction is removed (the lower restriction can’t be removed). Due to these restrictions, and the need of more accurate results, we chose to use GA for optimization. Let us consider the data given in Robert F. Dale and Aydin Öztürk [5] where we find 75 observations in ascending order with sample moments

$$\overline{x} = 5.109 \quad s = 1.013 \quad \hat{\alpha}_3 = 1.009 \quad \hat{\alpha}_4 = 3.344$$

By using the tables of Ramberg (1979)[1], and, Karian and Dudewicz[6], we can see that there is no possible combination of coefficients for skewness and kurtosis; there are no corresponding values for λ_3 *and* λ_4 . Now, using Algorithms 2009, a software built by David Moreno and Nelsón Fino[7], where we have the sample moments described above as inputs we found:

$$\hat{\lambda}_1 = 5.9674242470 \quad \hat{\lambda}_2 = 0.46737 \quad \hat{\lambda}_3 = 7.72384 \quad \hat{\lambda}_4 = 0.942217$$

with E 10^{-7} (method of moments).

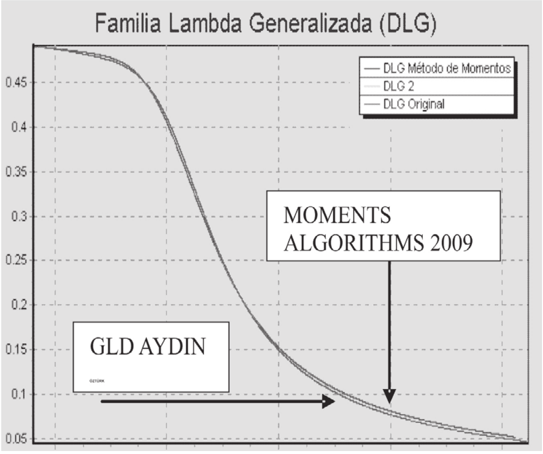


Figure 3. Probability function distributions of fitted samples.

We implemented a Kolmogorov-Smirnov test between the two solutions in order to see the goodness of fit. D = 0.0100334286, meaning (figure 4) that the distance between distribution functions is small and consequently it provides a good solution to the problem using the method of moments through GA. We also performed a hypothesis test where a value near to 1 means that the data comes from the same distribution. The result of this hypothesis was: 0.9999999403. Meaning that the values found for *GLD* $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ by Aydin Öztürk [5] and the ones found through Algorithms 2009[7] correspond to the same distribution. Using Algorithms 2009, we have also found values in the region “Solutions not available” from figure 2 using the method of moments [7].

Summary

A method for improving results for the GLD family based on Genetic Algorithm has been presented, the idea is to provide the needed tools for researchers to improve their investigations, this project came out as a need to automatize the search of the lambda values and as a way of improving the results obtained through the years using an algorithm that mimics the survival of the fittest. The methodology here proposed has found a solution with E 10^{-7} for the method of moments and solution points beyond the theoretical constraints imposed by [1] and [6].

References

[1] RAMBERG, J.S., TADIKAMALLA, P.R., DUDEWICZ, E.J., and MIKYTHA, E.F. (1979). "*A Probability Distribution And Its Uses In Fitting Data,*" Technometrics, 21, 201-214.

[2] GOLDBERG, DAVID. *Genetic Algorithms In Search Optimization And Machine Learning.* 1990.

[3] NUMERICAL RECIPIES IN C: THE ART OF SCIENTIFIC COMPUTING (ISBN 0-521-43108-5) Copyright (c) 1988-1992 by Cambrige University Press. http://www.nr.com

[4] http://lancet.mit.edu/ga/

[5] AYDIN ÖZTÜRK and ROBERT F. DALE (1985), "Least Squares Estimation Of The Parameters Of The Generalized Lambda Distribution", Technometrics, 27 No. 1, 81 - 84

[6] ZAVEN A. KARIAN and EDWARD J. DUDEWICZ. Fitting Statistical Distributions "*The Generalized Lambda Distributions and Generalized Bootstrap Methods*". CRC Press. Boca Raton London New York Washington, D.C. 2000.

[7] FINO RICARDO NELSON, MORENO DAVID LEONARDO. "Método Para Estimar Los Parámetros De La Distribución Lambda Generalizada Basado En Algoritmos Genéticos". Ingeniería de Sistemas, Universidad Antonio Nariño, Octubre de 2009.

[8] DUDEWICZ, E.J. and KARIAN, Z.A . "Fitting the Generalized Lambda Distribution To Data: a method based on percentiles" ACM Communications in Statistics, 28(3), 793-819 (1999).