Comunicaciones en Estadística
Estadística en la evaluación de la educación
Agosto 2016, Vol. 9, No. 2, pp. 255–270

Parameter estimation in mixture models using evolutive algorithms[1]

Estimación de parámetros en modelos de mezclas usando algoritmos evolutivos

Natalia Romero-Ríos[a]
natromerorio@unal.edu.co

Juan Carlos Correa[b]
jccorrea@unal.edu.co

Abstract

The mixture models are widely used in cases when there are elements that come from different populations, mixed in a superpopulation. There are traditional methods for the estimation of the parameters in mixture models: the Bayesian Method and the Expectation-Maximization (EM) algorithm. For that reason, in this work we propose the use of evolutive algorithms, such as genetic algorithms. We propose an algorithm for the comparison of evolutive and traditional methods, and we illustrate the use of this algorithm with a real application. We found that the evolutive algorithms are a competitive option to estimate the parameters in mixture models in the cases when the populations in the mixture follows a gamma distribution, the weights of the populations in the mixture are even and the sample size is bigger than 100 items. For the mixture of normal distributions and the estimation of the number of populations in a mixture, the traditional method is a better option than the genetic algorithm.

***Keywords*:** mixture estimation, mixture distribution, evolutive algorithms, genetic algorithms.

Resumen

Los modelos de mezclas son ampliamente usados en casos en los que se tienen elementos de poblaciones diversas, unidos en una super población. Hay métodos tradicionales para la estimación de los parámetros de modelos de mezclas, como lo son el bayesiano y el algoritmo de esperanza-maximización (EM). En esta investigación se propone usar los algoritmos evolutivos, como lo son los algoritmos

[1] Romero-Ríos, N., Correa, J.C. (2016) Parameter estimation in mixture models using evolutive algorithms. *Comunicaciones en Estadística*, **9**(2). 255-268.

[a] Universidad Nacional de Colombia, Sede Medellín, Colombia.

[b] Universidad Nacional de Colombia, Sede Medellín, Colombia.

genéticos, como método que puede servir para encontrar los parámetros de estimación de los modelos de mezclas. Para el desarrollo de este estudio se propone un algoritmo para la comparación de métodos evolutivos y tradicionales y se incluye un ejemplo de aplicación con datos reales. Se encontró que los algoritmos evolutivos son una opción competitiva para la estimación de parámetros en distribuciones de mezclas en los casos cuando las poblaciones en la mezcla siguen una distribución gamma, los pesos en las poblaciones son balanceados y el tamaño de muestra es mayor de 100 ítems. Para las mezclas de distribuciones normales y la estimación del número de poblaciones en una mezcla, el método tradicional es una mejor opción que el algoritmo genético.

***Palabras clave*:** estimación de mezclas, algoritmos evolutivos, algoritmos genéticos.

1 Introduction

The mixture models are statistical representations of an overall distribution with two or more subpopulations. The main idea behind these models is to represent the heterogeneity of the data (McLachlan & Basford 1988, Reynolds & Templin. 2004). In the real world, some populations are composed of subpopulations, for example, the height of a group of people is composed of at least two groups, women and men, but it can be generated of more than two groups, as the age or ethnicity. The mixture models can be represented as $x_1, \dots, x_n$ where each x_j is a p-dimensional vector, arising from a superpopulation G, which is a mixture of a finite number of populations, g, denoted as $G_1, \dots, G_g$ in some proportions $\pi_1, \dots, \pi_g$, respectively, where:

$$\sum_{i=1}^{g} \pi_i = 1, \qquad \pi_i > 0, \qquad (i = 1, \dots, g)$$

The probability density function (p.d.f) of x in G can be represented in the finite mixture form:

$$f(x;\phi) = \sum_{i=1}^{g} \pi_i f_i(x;\theta) \tag{1}$$

Where $f_i(x;\theta)$ is the p.d.f. of the G_i-th population, and θ denotes the vector of unknown parameters associated with the parametric forms adopted for the g densities. It is assumed that the vector $\phi = (\pi, \theta)'$ of unknown parameters belongs to some parameter space Ω.

For the modelling of the mixture models, or the estimation of the parameter ϕ and the number of population g, two approaches are widely used: the Bayesian approach and the classical approach. In the Bayesian approach the data is collected, plotted, smoothed and then, a given a prior distribution, as the Dirichlet distribution, is fitted to it. Further, the data is clustered and analysed (Crawford 1994). The computation of the posterior unlabeled observation is difficult due to the form of the likelihood, because the number of terms grows exponentially with the sample size n, and generally cannot be solved using analytical methods (West 1993). The classical approach is the maximum likelihood estimation (McLachlan & Basford 1988). The likelihood estimation uses the EM algorithm (E for expectation, and M for maximation) of Dempster, Laird and Rubin (McLachlan & Basford 1988). To run this algorithm, it is needed some starting values for $\phi, \phi^{(0)}$ on the equation 1, or to initially partition the data into the specified number of groups g, and take $\phi^{(0)}$, as the estimate of ϕ based on this partition, as it represented the true grouping of the data (McLachlan & Basford 1988, Susko et al. 1998, Gallegos & Ritter 2009, Snee 1973).

We propose the use of evolutive algorithms to estimate the parameters of mixture distributions. Evolutive algorithms are methods of stochastic search, that can work in very complex problems without the assumptions of the traditional methods, such as the continuity and the existence of derivatives (Haupt & Haupt 2004). Some examples are Simulated Annealing, Taboo Gearch and Genetic Algorithms (Fouskakis & Draper 2002). The first one, Simulated Annealing, works as an analogy of the change of temperature of the materials under an annealing process (Metropolis et al. 1953). Taboo Search uses a structured method to find the maximum of a function avoding local optima by imposing restrictions or ”taboo’ and searching on the entire parameter space (Glover 1989). Finally, Genetic Algorithms uses biological concepts as evolution, crossbreeding and selection to find the maximum of a function (Zhu & Chipman 2006, Scrucca 2013, Denning 1992).

1.1 Algorithm

Genetic algorithms (GA) are stochastic searches models (Zhu & Chipman 2006) first proposed by Holland (1975) in (Fouskakis & Draper 2002). These models work as an analogy to Darwinian evolution, with their structural blocks, chromosomes, and making those evolve by selection, crossover and mutation (Denning 1992). The innovations proposed were “using bit string representations, proportional selection and crossover as the main operators” (Scrucca 2013). To implement a GA, first we must know the function to be optimized; later, a set of n chromosomes of length p are generated at random. The next step is to evaluate the fitness for every chromosome, and to arrange them by pairs, making the most fitted more likely to crossover, and there is a chance to their offspring to mutate. Later, only the most fitted between parents and their offspring are allowed to continue, and new chromosomes are generated.

Two generic algorithms were used in this work. Both were made on `R` (R Core Team 2014). One was used for the known number of distributions and it is described on figure 1, because we can compute the error on the estimation for every parameter, and another when the number of populations was unknown, figure 2, because it needs a different approach.

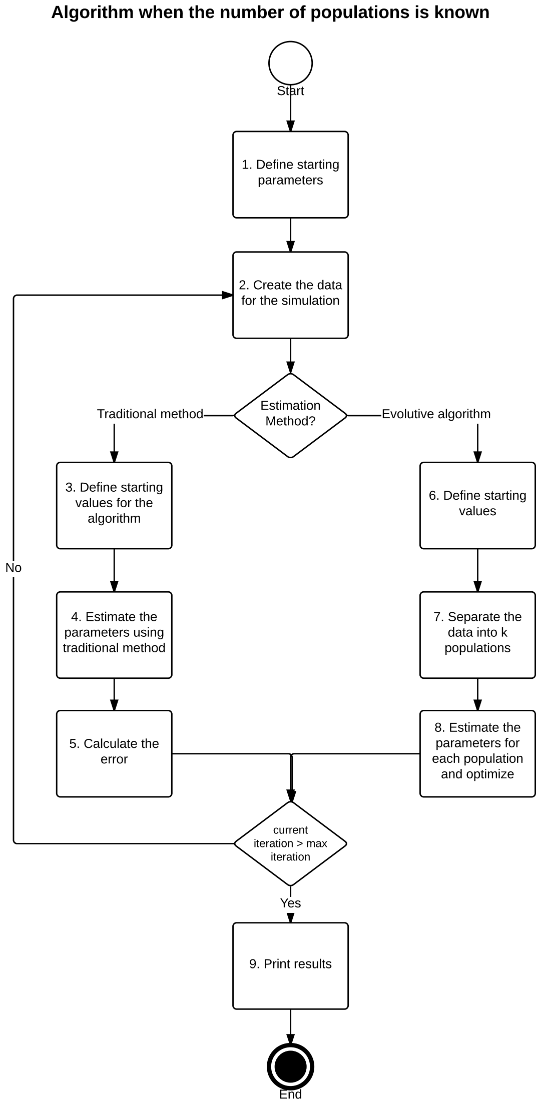
Algorithm when the number of populations is known
Start
1. Define starting parameters
2. Create the data for the simulation
Estimation Method?
Traditional method
Evolutive algorithm
3. Define starting values for the algorithm
4. Estimate the parameters using traditional method
5. Calculate the error
6. Define starting values
7. Separate the data into k populations
8. Estimate the parameters for each population and optimize
current iteration > max iteration
No
Yes
9. Print results
End

Figure 1: *Algorithm when the number of populations is known. Source: elaborated by authors.*

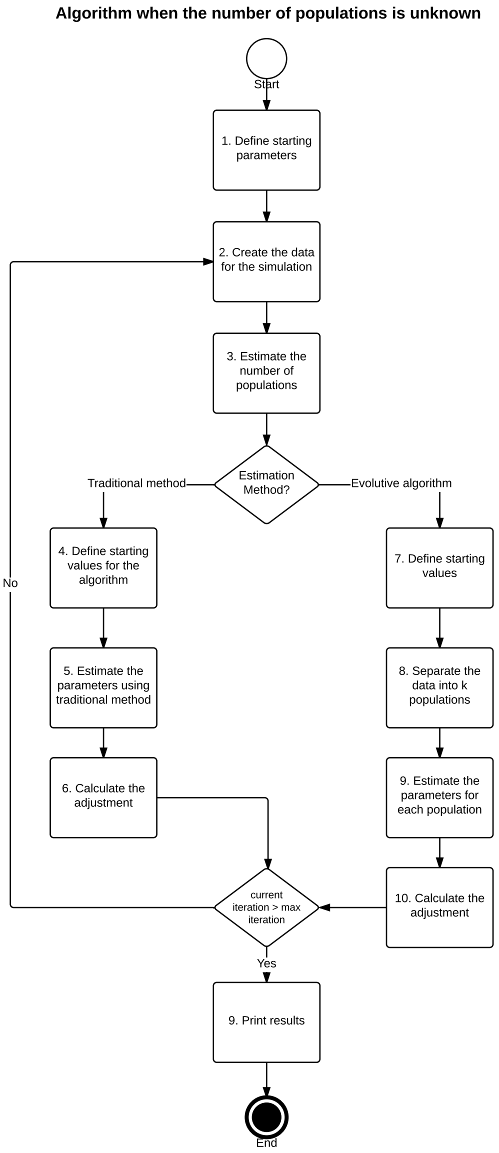
Algorithm when the number of populations is unknown
Start
1. Define starting parameters
2. Create the data for the simulation
3. Estimate the number of populations
Estimation Method?
Traditional method
Evolutive algorithm
4. Define starting values for the algorithm
5. Estimate the parameters using traditional method
6. Calculate the adjustment
7. Define starting values
8. Separate the data into k populations
9. Estimate the parameters for each population
10. Calculate the adjustment
current iteration > max iteration
No
Yes
9. Print results
End

Figure 2: *Algorithm when the number of populations is unknown. Source: elaborated by authors.*

2 Simulation Results

To make the comparation between traditional methods and evolutive algorithms to estimate the parameters in mixture models, we implement a simulation study. This allow us to know the real parameters and find the error within the estimation. The basic form for a mixture functions follow the equation 1. For this case, we are going to use a mixture of two normal distributions with parameters $\theta_1 = (\mu_1, \sigma_1)$ and $\theta_2 = (\mu_2, \sigma_2)$, respectively so the equation 1 follows the form:

$$f(x; \mu_1, \sigma_1, \mu_2, \sigma_2, \pi) = \pi * f_1(x, \mu_1, \sigma_1) + (1 - \pi) * f_2(x, \mu_2, \sigma_2)$$

In this case, the values to estimate are:

μ_1, σ_1 Parameters of the first population

μ_2, σ_2 Parameters of the second population

π Population weights

The configuration for the simulations is shown in table 1.

Table 1: *Values for the parameters of the simulation study. Source: elaborated by authors.*

Factor	**Levels**			
Populations	Known	Unknown		
Mixture of distributions	Normal	Gamma		
k Number of populations	2	3	5	
Separation between means	2	3	5	
π_i Population weights	5	10	25	50
Population size	30	50	100	200

To compute the distance between the estimated and the real density, the Hellinger distance (HD) is used as an approach to measure the distance between the true and estimated densities, the one with the true parameters used for the simulation, $f(x)$, and the one with the parameters given by the EM and GA, $g(x)$ (Beran 1977). This is shown in equation (2). This estimator has been used before in mixtures of parametric families, as described by (Wu & Karunamuni 2009) and (Adele & Cordero-Braña 1996) and it has been shown that this estimator is robust. To analyze the result, when the distance is zero, it means that the estimated values are the same as the real ones, for this reason, the best method is the one to achieve the minimum values (Adele & Cordero-Braña 1996).

$$HD = \int_{-\infty}^{\infty} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx \approx \sum_{i=1}^{M} \left(\sqrt{f(x_i)} - \sqrt{g(x_i)}\right)^2 \qquad (2)$$

Where X is a variable created to estimate the approximated distance expressed in equation 2 in the interval $X \in I$, $I = \{(\mu_1 - 3\sigma_1);(\mu_k + 3\sigma_k))\}$, for the normal mixture, and $I = \{(\alpha_1 - 3\beta_1);(\alpha_k + 3\beta_k))\}$ for the gamma mixture. x_i, $i = 1...500$, being $M = 500$, estimated in a grid of X, and the population k is the population with the mean or the scale parameter for the normal and gamma cases, respectively. $f(x)$ is the real density, with the parameters used to generate the data in the simulation. $g(x)$ is the estimated density, with the parameters obtained with the EM or the GA.

The simulations were run as shown in table 1, and the results for the experiments when there are 5 known populations are shown in figures 3 and 4 for the mixture of normal populations and gamma populations, respectively. The name codes are as follows:

Weight: Is the percentage weight of the first distribution, the weight of the other distributions in the mixture were asigned in equal parts to complete the 100%

GA: Genetic Algorithm

EM: Expectation Maximization Algorithm

Mean: Is the mean between the results of the simulations

Hellinger distance: Is the Hellinger distance calculated with equation (2)

2.1 Results of the estimation of the parameters when the number of populations is known

For the mixture of normal populations when the number of populations is known, with the case for 5 populations found in figure 3 and figure 4 for the normal and gamma mixture, respectively, the remarkable findings are: When the weights of the populations in the mixture were even, the Hellinger distance was smaller than when the weights were uneven, and when the weight of one of the populations was smaller than 10%, the GA could not detect the population, and could not estimate the parameters. That estimation was not a problem for the EM algorithm. In general, as the sample size increases, the Hellinger distance decreases. The number of populations in the mixture increases the calculated Hellinger distance, for both methods, GA and EM. The separation between the means of each population in the mixture increases the mean Hellinger distance, for the GA results.

When the number of populations is known, in a mixture of normal populations, the EM algorithm had better results than the GA, and in all cases could estimate the parameters, even when the sample size and the weight were small, but the estimation of the parameters of the GA were similar.

In a mixture of gamma populations, the EM algorithm had numerical inestability with the estimation of the parameters, but the GA could yield results in all the

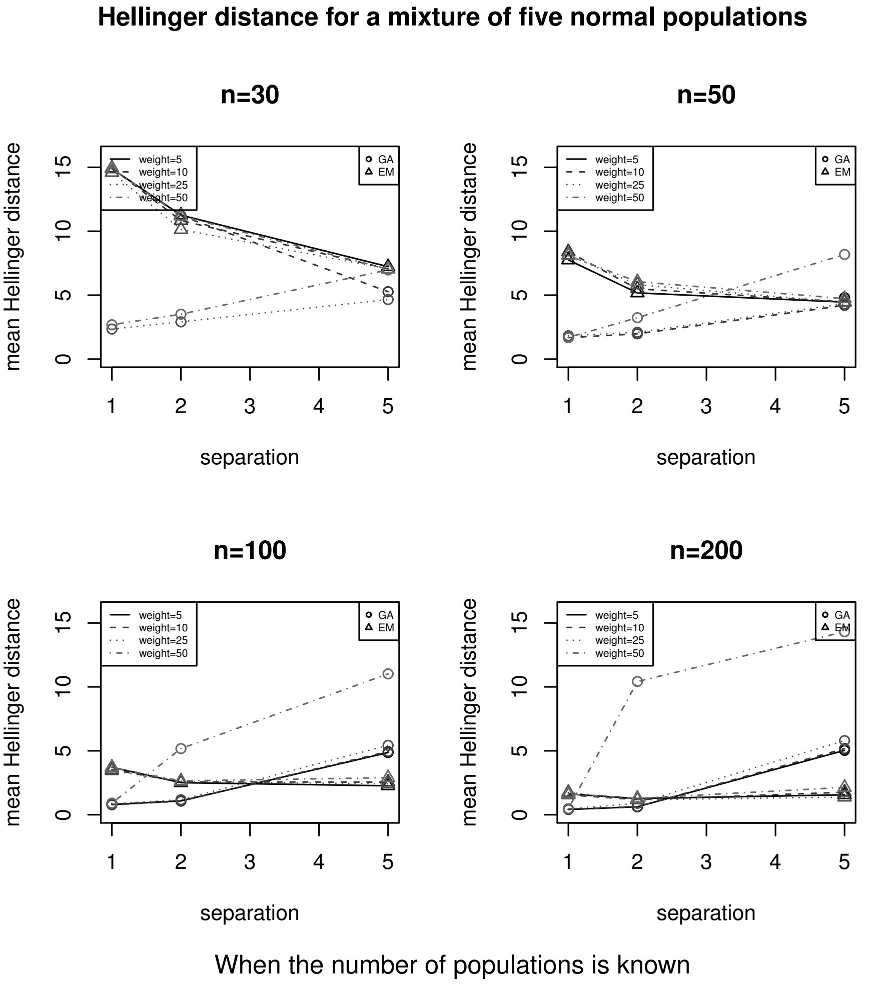

Figure 3: *Plot of the Hellinger distance for the mixture of 5 normal populations, in a mixture with known number of populations. Source: elaborated by the authors.*

cases, except for the ones when the weight of one population and the sample size were small. Also, in general, the GA had better results of the EM in this case, with smaller Hellinger distance.

2.2 Results of the estimation of the number of populations

For the estimation of the number of populations, the inicial number of populations was set as 3 + *real number of populations* for both methods: GA and EM, the simulations were made with the same factors described in table 1 and the results

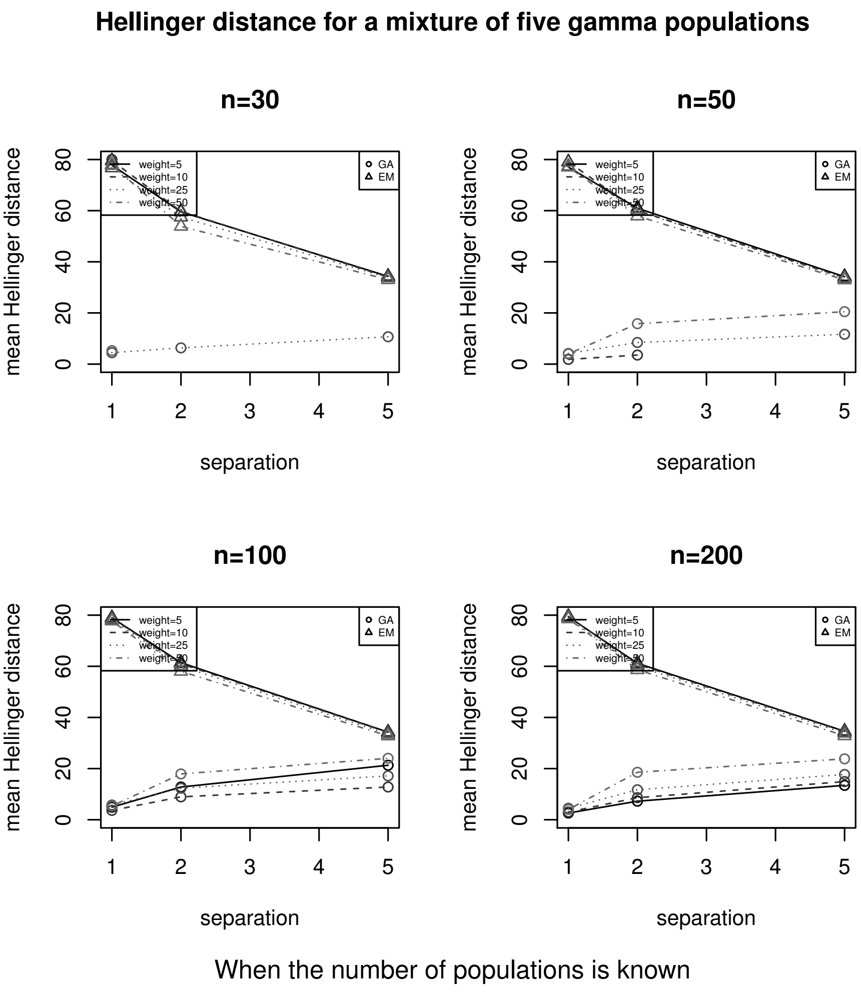

Figure 4: *Plot of the Hellinger distance for the mixture of 5 gamma populations, in a mixture with known number of populations. Source: elaborated by the authors.*

can be seen in figures 5 and 6, for the case with 5 populations, respectively.

It can be seen that for the mixture of normal populations, in figure 5, both methods diverge in their behavior, because the GA overestimates and the EM underestimates the number of populations. For the GA the closest experiments were the ones with small sample data, this could be for the same reason when the number of populations was known, when they could not estimate all the data and yielded NaN.When the sample size was big enough, 100 or 200 the method in all the cases estimated the inicial set of populations (8). For the EM algorithm, all the results were really close, but the best results were obtained when the identification was

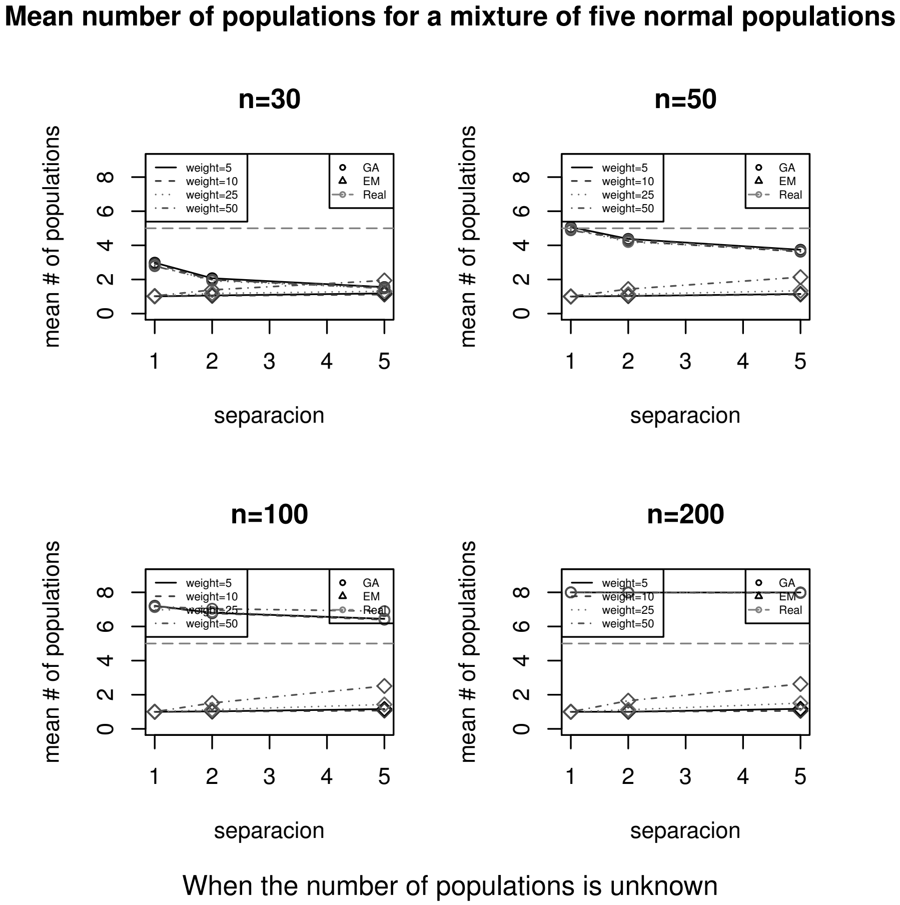

Figure 5: *Plot of the mean estimated number of pofbpulations for the mixture of 5 normal populations, in a mixture with unknown number of populations. Source: elaborated by the authors.*

easy, as exposed in the previous analysis, when the sample size, separation and weight was large. In the other cases, the method could not find all the populations, also, this method had a smaller variation compared with the GA.For this reason, the conclusion is that neither of the methods, GA nor EM algorithm are exact, neither of both methods in neither scenario had an exact performance, and this can be checked by looking carefuly the images.

This behaviour, for the EM algorithm, is repeated in the mixture of gamma populations and it can be seen in figure 6. The GA could no estimate the number of population, because of the tendency to overestimate the number of populations to the initial parameter of 8 populations, this lead to numerical inestability and

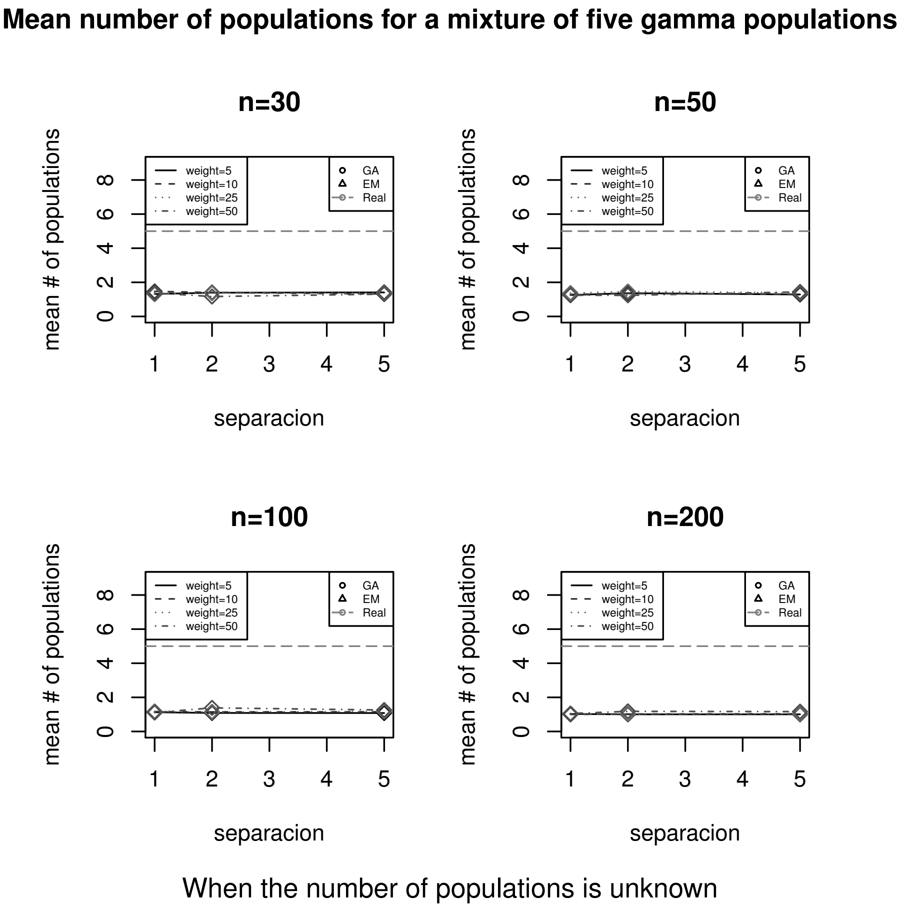

Figure 6: *Plot of the mean estimated number of populations for the mixture of 5 gamma populations, in a mixture with unknown number of populations. Source: elaborated by the authors.*

posterior errors. For this reason, for the mixture of gammas, the EM is better in the estimation of the number of populations.

2.3 Illustrative example

The data for this illustration was taken from a study conducted by Estrada et al. (1988), the Instituto del Seguro Social gave the permission to use de data set. This study had as an objective to measure 69 anthropometric parameters from a working population in Colombia. The data was taken from males and females from 20 to 60 years old, and the aim was to get a characterization of the population, and with the information taken from this database to get to design spaces and equipment for the use of the Colombian workers, because historically these have been designed using international standards or heuristically. From this study, the data on BMI (Body Mass Index) has been selected as the variable to analyze, because of the importance to describe the body and therefore the designs to do for the colombian workers, also is a variable that is important to show the risk of mortality by circulatory diseases or cancer (Estrada et al. 1988). The histogram and the density can be seen in figure 7 where it shows a form of a couple, but with a heavy tail on the right, and a little hump around a BMI of 30. Looking very carefully, it can be seen another humps on 24 and 28. The Kolmogorov Smirnov test was made to check normality, for a two sided hypotesis. We observed a $p - value < 2.20E - 16$, and this analysis confirms that the distribution is not a normal one. For this reason, an analysis using a mixture of distributions is appropriate.

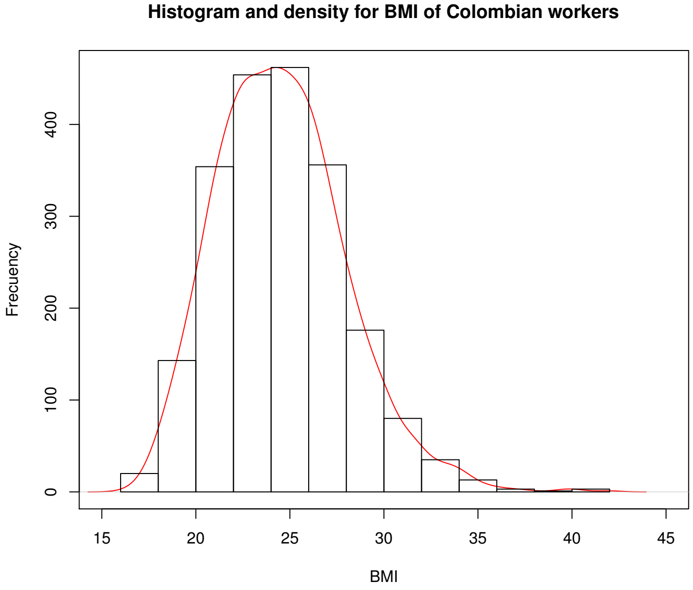

Figure 7: *Histogram and estimated density for data about the BMI of colombian workers. Source: elaborated by authors.*

The number of population was made using the EM algorith, and gave as a result 4 populations in the mixture, so the parameter comparison was with the GA and EM algorithm with 2, 3 and 4 estimated populations. The parameters estimated are shown in table 2. The parameters of every populations were different from the calculations of both methods, and as a way to assess the adjustment to the data, the graph 8 was created. In this graphic, the best method for this data set is the EM algorithm with 3 populations, because it is the one that best fits the estimated density.This method gives the information to conclude that there are three groups of Colombian workers, one with a the 24% of the people, with a healthy BMI, with mean 21, the majority 64% with overweight with a BMI of 25, and with a standard deviation of 2,6, and the last one with a 12% of people, with a BMI of 28, close to the obesity.

Table 2: *Parameters estimated from the mixture of BMI of colombian workers. Source: elaborated by authors.*

Popula.	Pop number	GA			EM		
		π	μ	σ	π	μ	σ
2	1	0.4567	26.3176	3.6591	0.7938	23.7285	2.7923
	2	0.5433	23.0441	2.3605	0.2062	27.6599	3.8662
3	1	0.2657	23.0418	2.3409	0.2381	21.1638	1.6954
	2	0.4681	25.9696	3.9413	0.1191	28.7908	3.9641
	3	0.2662	23.5178	2.1600	0.6429	25.0014	2.5691
4	1	0.1590	23.2227	2.6489	0.2934	21.3270	1.7535
	2	0.3252	26.9433	3.7198	0.0065	24.0586	0.0294
	3	0.3662	23.1989	2.4640	0.5700	25.2685	2.4292
	4	0.1495	23.9916	2.7161	0.1302	28.6128	3.9589

As a conclusion, the methods can be used for real case studies with results that can describe the data. As a recommendation, we endorse further studies of the number of populations, because it is a critical input and the methods here exposed are not very accurate for the estimation of the number of populations in the mixture. We recommend to follow the EM algorithm for the estimation of the number of populations, and next using an evolutive algorithm if the distribution is not a mixture of normal populations.

3 Conclusions

In this study, a comparison between traditional methods and evolutive algorithms was made to estimate the parameters in mixture models, with mixtures composed of normal and gamma populations. The factors of: type of mixture, number of populations, population weight, sample size and separation between means was made. The objective was to estimate the parameters of the mixture in the case when the number of populations is known, and estimate the number of populations, when this number is unknown. For this comparison, a software in *R* (R Core Team 2014) was developed and it is proposed in this study.

Of this comparison, we can conclude that when the number of populations is known, in a mixture of normal populations, the EM algorithm had better results than the GA, because the Hellinger distance was smaller and in all cases could estimate the parameters, even when the sample size and the weight were small. In the case of a mixture of gamma populations, the EM algorithm had numerical

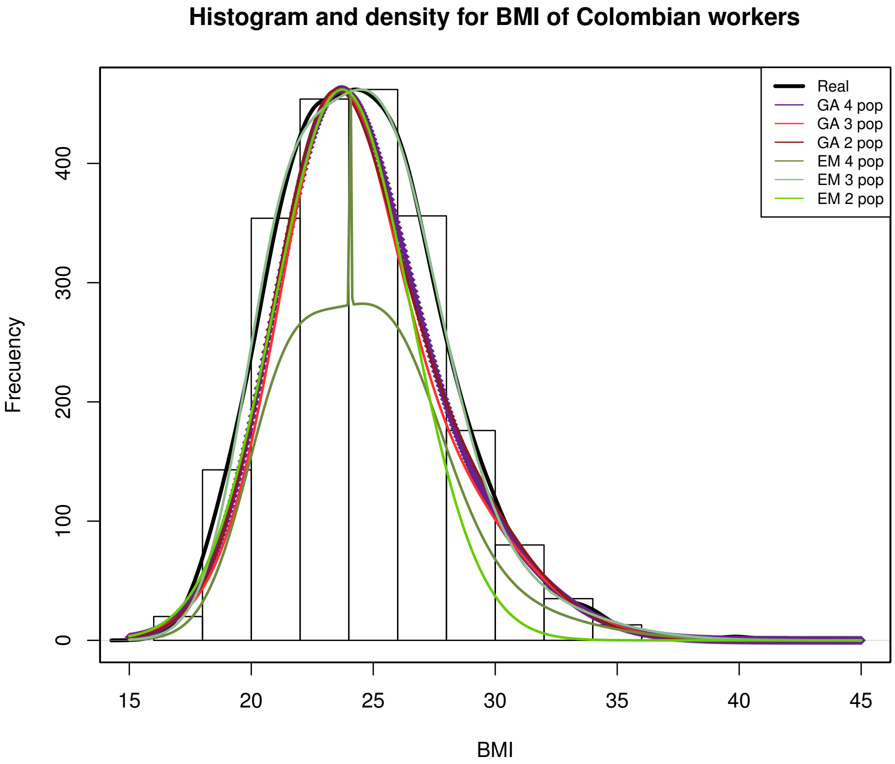

Figure 8: *Comparison of method to estimate the mixture of BMI of colombian workers. Source: elaborated by authors.*

inestability with the estimation of the parameters, but the GA could yield results in all the cases, except for the ones when the weight of one population and the sample size were small. Also, in general, the GA had better results of the EM in this case, with smaller Hellinger distance. For this reason, evolutive algorithms are a competitive option to traditional methods to estimate the parameters, when the populations in a mixture are not normal.

For the case when the number of populations is unknown the GA overestimates the number of populations, and it yields as result, the initial parameter of number of populations, for this reason, the EM is a better option for the estimaton of the number of populations in a mixture.

Recibido: 10 de febrero de 2016
Aceptado: 11 de abril de 2016

References

Adele, C. & Cordero-Braña, O. I. (1996), 'Minimum hellinger distance estimation for finite mixture models', *Journal of the American Statistical Association* **91**(436), 1716–1723.
*http://www.jstor.org/stable/2291601

Beran, R. (1977), 'Minimum hellinger distance estimates for parametric models', *The Annals of Statistics* **5**(3), 445–463.
*http://www.jstor.org/stable/2958896

Crawford, S. L. (1994), 'An application of the laplace method to finite mixture distributions', *Journal of the American Statistical Association* **89**(425), 259–267.
*http://www.jstor.org/stable/2291222

Denning, P. (1992), 'The science of computing: Genetic algorithms.', *American Scientist* **80**(1), 12–14.

Estrada, J., Camacho, J., Restrepo, M. & Parra, C. (1988), 'Parámetros antropométricos de la población laboral colombiana 1995 (acopla95)', *Revista Facultad Nacional de Salud Pública* **15**(2), 112–139.

Fouskakis, D. & Draper, D. (2002), 'Stochastic optimization: a review.', *International Statistical Review / Revue Internationale de Statistique* **70**(3), 315–349.

Gallegos, M. & Ritter, G. (2009), 'Trimmed ml estimation of contaminated mixtures', *Sankhya: The Indian Journal of Statistics, Series A* **71**(2), 164–220.

Glover, F. (1989), 'Tabu Search Part I', *ORSA Journal on Computing* **1**(3), 190–206.

Haupt, R. & Haupt, S. (2004), *Practical Genetic Algorithms*, Wiley.

McLachlan, G. & Basford, K. (1988), *Mixture models: inference and applications to clustering*, Marcel Dekker.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953), 'Equation of state calculation by fast computing machines.', *Journal of Chemical Physics* **21**(6), 1087–1091.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org

Reynolds, J. & Templin., W. (2004), 'Comparing mixture estimates by parametric bootstrapping likelihood ratios.', *Journal of Agricultural, Biological, and Environmental Statistics* **9**(1), 54–74.

Scrucca, L. (2013), 'Ga: A package for genetic algorithms in r.', *Journal of Statistical Software* **53**(4), 1–37.

Snee, R. (1973), 'Techniques for the analysis of mixture data.', *Technometrics* **15**(3), 517–528.

Susko, E., Kalbfleisch, J. D. & Chen, J. (1998), 'Constrained nonparametric maximum-likelihood estimation for mixture models.', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **26**(4), 601–617.

West, M. (1993), 'Approximating posterior distributions by mixture.', *Journal of the Royal Statistical Society. Series B (Methodological)* **55**(2), 409–422.

Wu, J. & Karunamuni, R. J. (2009), 'On minimum hellinger distance estimation', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **37**(4), 514–533.
*http://www.jstor.org/stable/25653496

Zhu, M. & Chipman, H. (2006), 'Darwinian evolution in parallel universes: A parallel genetic algorithm for variable selection.', *Technometrics* **48**(4), 491–502.