L JAST
Libya Journal Applied For Science and Technology
مجلة ليبيا للعلوم التطبيقية والتقنية

Volume 5.Issue1
December2018

Fitting a finite mixture of exponential distribution to data for the transfer size (in bytes) of documents returned to requesting clients from the World –Wide–Web using Libyan Internet proxy server

Adel Ali Ewhida
Tripoli University, Faculty of Science, Department of Statistics, P.O. Box 13219, Tripoli, Libya

Abstract

We try to fit finite mixtures of exponential distributions to data for the transfer size (in bytes) of documents returned to requesting clients from the World –Wide–Web using Libyan Internet proxy server. We use two algorithms to fit a finite mixture of exponential distributions directly to the data. In this case, we use the three-moment matching algorithm [12] to fit a mixture of two exponential distributions to the first three moments obtained from the data, and we use the recursive fitting algorithm [4] to fit a finite mixture of exponential distributions to the empirical $ccdf$ obtained from the data.

Words: Mixtures of exponential distributions, Likelihood Estimation.

1. Introduction

When a population is composed of several distinct subpopulations, the distribution of a variable defined on this population is a finite mixture distribution. Mixture distributions have received increasing attention in the statistical literature mostly because of the considerable number of areas in which such distributions are encountered. An important example is fisheries length-frequency analysis, where a fish population is a mixture of distinct age groups. Other examples for fitting mixture distributions come from medical research, biology and sedimentology (e.g., Macdonald [7], Everitt and Hand [3], Titterington, Smith and Makov [10]).

The statistical analysis of data from a population that has a finite mixture distribution has proven not to be straightforward. In particular, there is no generally simple formula for estimating the unknown parameters since this is often a nonlinear estimation problem. This and other problems that occurred with mixture distributions have been studied by many researchers mostly in the area of finite mixtures of normal distributions, but there are also some references on direct applications of the mixtures of gamma, exponential, beta, binomial or poisson distributions (Titterington, Smith and Makov [10]).

In particular, there are three important studies on mixture distributions presented by Everitt and Hand [3], Titterington, Smith and Makov [10], and McLachlan and Basford [8]. There are also review articles on mixture distributions by Holgersson and Jorner [6], Gupta and Huang [5], and Redner and Walker [9].

This paper focus on fitting finite mixtures of exponential distributions to data for the transfer size (in bytes) of documents returned to requesting clients from the World –Wide–Web using Libyan Internet proxy server.

2. Material and Methods

A-The Three-moment Matching Algorithm

Whitt [12] showed that there is an H_2 distribution (a mixture of two exponential distributions) with the same μ_1, μ_2 and μ_3, the first three moments of the data, if and only if $\mu_1 \geq 0$, $\mu_2\mu_1^{-2} \geq 2$ and $\mu_3\mu_1 \geq 1.5\mu_2^2$. If μ_3 is small, it can be replaced by something slightly larger than $\frac{1.5\mu_2^2}{\mu_1}$.

In other words, given the first three moments μ_1, μ_2 and μ_3 which satisfy the previous conditions, there is an H_2 distribution with the same first three moments. The H_2 parameters $\lambda_1, \lambda_2, p_1$ and p_2 can be obtained by

$$\lambda_i^{-1} = \left[(x + 1.5y^2 + 3\mu_1^2) \pm \sqrt{(x + 1.5y^2 + 3\mu_1^2 y)^2 - 12\mu_1^2 xy}\right]\left(\frac{1}{6\mu_1 y}\right) \geq 0 \qquad (1)$$

$$p_1 = \left(\frac{\mu_1 - \lambda_2^{-1}}{\lambda_1^{-1} - \lambda_2^{-1}}\right) \geq 0 \qquad (2)$$

And

$$p_2 = 1 - p_1 \geq 0 \qquad (3)$$

Where

$i = 1,2, x = \mu_1\mu_3 - 1.5\mu_2^2, y = \mu_2 - 2\mu_1^2$

and $x, y \geq 0$; see Whitt [11] and Abate and Whitt [1].

B-The Recursive Fitting Algorithm [4]

The algorithm as follows: first choose r, the number of exponential components, and c_i arguments where we will match quintiles such that $0 < c_r < c_{r-1} < \cdots < c_1$. We choose a constant b such that $1 < b < \frac{c_i}{c_i+1}$ for $i = 1, \ldots, r$. Then the approximate and values of the parameters λ_1 and p_1 obtained by letting $p_1 e^{-\lambda_1 c_1} = F_{U_1}^c(c_1)$ and $p_1 e^{-\lambda_1 b c_1} = F_{U_1}^c(bc_1)$ are

$$\lambda_1 = \frac{1}{(b-1)c_1} \ln\left[\frac{F_{U_1}^c(c_1)}{F_{U_1}^c(bc_1)}\right] \qquad (4)$$

$$p_1 = F_{U_1}^c(c_1) e^{\lambda_1 c_1} \qquad (5)$$

and for $2 \leq i < r$ the approximate values of the parameters λ_i and p_i obtained by letting $p_i e^{-\lambda_i c_i} = F_{U_i}^c(c_i)$ and $p_i e^{-\lambda_i b c_i} = F_{U_i}^c(bc_i)$ are

$$\lambda_i = \frac{1}{(b-1)c_i} \ln\left[\frac{F_{U_i}^c(c_i)}{F_{U_i}^c(bc_i)}\right] \qquad (6)$$

$$p_i = F_{U_i}^c(c_i) e^{\lambda_i c_i} \qquad (7)$$

Finally, the approximate value of the parameter p_r is

$$p_r = 1 - \sum_{j=1}^{r-1} p_j \qquad (8)$$

and the approximate value of the parameter λ_r obtained by letting $p_r e^{-\lambda_r c_r} = F_{U_r}^c(c_r)$ is

$$\lambda_r = \frac{1}{c_r} \ln\left[\frac{p_r}{F_{U_r}^c(c_r)}\right] \qquad (9)$$

where $F_{U_1}^c(u) = F_U^c(u)$, and for $2 \leq i < r$,

$$F_{U_i}^c(u) = F_{U_{i-1}}^c(u) - \sum_{j=1}^{i-1} p_j e^{-\lambda_j u} \qquad (10)$$

3. Software

All calculations were performed with software developed by me.

L JAST
Libya Journal Applied For Science and Technology
مجلة ليبيا للعلوم التطبيقية والتقنية
Volume 5.Issue1
December2018

4. Data and Results

We consider a random sample of size 2277 from the transfer size (in bytes) of documents returned to requesting clients from the World –Wide–Web using the Libyan Internet proxy server, the data was collected in 7 days between November 1 and December 28 in 2017, for which we try to fit a finite mixture of exponential distribution. Firstly, we consider fitting an H_2 distribution to the data using the three-moment matching algorithm. The parameters of the fitted H_2 distribution are

$p_1 = 0.0003$, $p_2 = 0.9997$,

$\lambda_1 = 2.756 \times 10^{-7}$ and $\lambda_2 = 1.4 \times 10^{-4}$.

Plot (a) in Figure 1 shows that the approximating H_2 $ccdf$ fit poorly to the empirical $ccdf$ of the data. Figure 2 (a) and (b) shows that the fitted H_2 distribution had very large absolute and relative errors. Moreover, using the Cramer-Von Mises goodness-of-fit modified test we reject the hypothesis $H_0 : X \sim H_2$ at the 0.1 significance level,

since.

$$(\ \widetilde{CM} - \frac{0.4}{n} + \frac{0.6}{n^2})(1 + \frac{1}{n}) = 12.0463 > 0.347$$

Secondly, we consider fitting an H_4 distribution (a mixture of 4 exponential distributions) to the empirical $ccdf$ of the data using the recursive fitting algorithm. The parameters of the fitted H_4 distribution are given in Table 1.

Plot (b) Figure 1 shows that the approximating H_4 ccdf also fit poorly to the empirical ccdf of the data. Figure 2 (c) and (d) show also that the fitted H_4 distribution had very large absolute and relative errors. Table 2 shows that the first three moments of the H_4 distribution are not very close to the first three moments of the data. Moreover, using the Cramér-Von Mises goodness-of-fit modified test we reject the hypothesis $H_0 : X \sim H_4$ at the 0.1 significance level, since .

$$(\ \widetilde{CM} - \frac{0.4}{n} + \frac{0.6}{n^2})(1 + \frac{1}{n}) = 3.2348 > 0.347$$

Thus a new fit with more exponential components should be considered to approximate the distribution of this data (a transfer size data). However, it was impossible to fit more than 4 exponential components directly to the data because the range of the empirical $ccdf$ is limited.

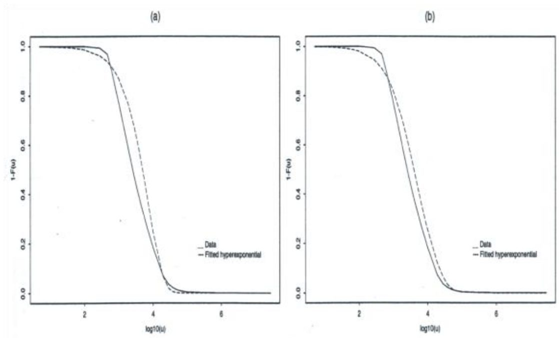

Figure 1: Comparison between the $ccdf$ of the data with two finite mixtures of exponential distributions, in plot (a) using the three moment matching algorithm for the H_2 fit, plot (b) using the recursive fitting algorithm for the H_4 with $c_1 = 10^5$ and $c_4 = 10^3$.

L JAST
Libya Journal Applied For Science and Technology
مجلة ليبيا للعلوم التطبيقية والتقنية
Volume 5.Issue1
December2018

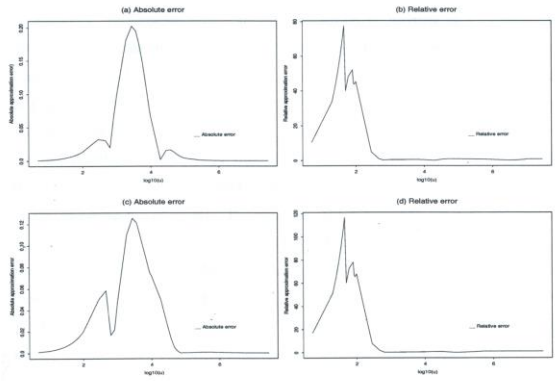

Figure 2: A comparison between the absolute and relative errors of two fitted mixtures of exponential distributions to the data, in plot (a) and (b) using the three-moment matching algorithm for the H_2 fit, plot (c) and (b) using the recursive fitting algorithm for the H_4 fit with $c_1 = 10^5$ and $c_4 = 10^3$.

Table 1:Parameters of the H_4 fit to the transfer size data.

i	p_i	λ_i
1	9.369×10^{-3}	9.808×10^{-6}
2	0.3017	6.657×10^{-5}
3	0.2597	1.172×10^{-4}
4	0.4292	3.698×10^{-4}

Table 2: Moments of the data sample and the fitted H_4 distribution.

Moment	*Data*	H_4 Distribution
1	$8,24 \times 10^3$	8.864×10^3
2	$8,076 \times 10^9$	3.75×10^8
3	$8,68 \times 10^{16}$	6.673×10^{13}

5. Conclusion

Applying both algorithms directly to the data indicates that both of them give poor mixtures of exponential distributions fits, whereas the second algorithm gives a better fit to the data if we first fit a heavy tail distribution to data and then apply the second algorithm to the fitted heavy tail distribution (Ewhida [2]). Therefore, depending on the application, either algorithm may be preferable.

References

[1] Abate, J. and Whitt, W. (1987). Transient Behavior of regulated Brownian motion, I:Starting at the origin. Adv. Aplpl. Prob., 19, 560-598.

[2] Ewhida, A. (2018). The weibull distribution as mixture of exponential distributions. Al-Tarbawi Journal, Al-Khoms. Libya, No. 13, July 2018, pp 360-367.

[3] Everitt, B.S. and Hand, D.J. (1981). Finite Mixture Distributions. Chapman and Hall, London.

[4] Feldmann, A. and Whitt, W. (1996). Fitting mixtures of exponentials to long tail distributions to analyze network performance models. AT & T Research.

[5] Gupta, S.S. and Huang, W.T. (1981). On mixtures distributions: a survey and some new results on ranking and selection. Sankhy , B 43, 245-290.

[6] Holgersson, M. and Jorner, U. (1978). Decomposition of a mixture into normal components: a review. Int. J. Bio.-Med. Comput., 9, 367-392.

[7] Macdonald, P.D.M. (1991). Practical aspects of fitting mixture distribution to data.

[8] McLachlan, G.J. and Basford, K.E. (1988). Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

[9] Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, 26, 195-239.

[10] Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). Statistical Analysis of Finite Mixture Distributions. New York: Wiley.

[11] Whitt, W. (1979). Approximating a Point Process by a renewal Process: A General Framework. Bell Laboratories.

[12] Whitt, W. (1982). Approximating a Point Process by a renewal Process, I: two basic methods. Operations Research, 30, 125-147.