Journal of Modern Applied Statistical Methods

Volume 10 | Issue 2 Article 17

11-1-2011

Maximum Log Likelihood Estimation using EM Algorithm and Partition Maximum Log Likelihood Estimation for Mixtures of Generalized Lambda Distributions

Steve Su
University of Western Australia, allegro.su@gmail.com

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 599-606

1538 – 9472/11/$95.00

Maximum Log Likelihood Estimation using EM Algorithm and Partition Maximum Log Likelihood Estimation for Mixtures of Generalized Lambda Distributions

Steve Su
University of Western Australia,
Perth, Australia
Covance Pty Ltd, Sydney, Australia

Two mixture distribution fitting methods based on maximizing the likelihood using generalized lambda distributions are presented. The fitting algorithms are demonstrated on various data and the strengths and weakness of the algorithms which can influence their use under different mixture modeling situations are discussed. The procedures described are available in GLDEX package in R.

Key words: Fitting distributions, prior distributions, empirical data analysis, mixture distributions, generalized lambda distributions.

Introduction

Mixture distribution modeling is a substantial area of interest among statisticians; many works regarding fitting mixtures have appeared in the literature. Böhning and Seidel (2003) discussed the general strategy used in confronting various problems associated with mixture distribution modeling. Although there are generic works, such as finding initial values to ensure better optimization of the mixture fitting scheme (Karlis & Xekalaki, 2003) and finding the optimal number of components of mixtures (Miloslavsky & van der Laan, 2003), no work has been presented on using mixtures of the generalized Lambda distributions to fit multi-modal data. This is an important development because the use of generalized Lambda distributions has advantages over traditional distributions such as Normal, Weibull and Exponential in the sense that they have overwhelmingly rich shapes and can handle a wide range of different data sets (Freimer, et al., 1988; Karian & Dudewicz, 2000; Okur, 1988; Su, 2010a, 2010b, 2005, 2007a, 2007b). Fitting a mixture of generalized Lambda distributions can therefore be very beneficial because it is much more efficient to fit distributions to data using a smaller range of distributions rather than choosing and comparing across a wide range of different combination of distributions.

Though generalized Lambda distributions are flexible their uses are not as widespread; this may be due to the fact that these distributions are only explicitly defined by quantiles, thus, extensive numerical methods are required to perform standard calculations, such as finding the probability under the curve. As computing power continues to grow, maximum likelihood estimations conducted numerically may become more popular. This article discusses two different ways of fitting mixtures using generalized Lambda distributions (GλDs).

Methodology

The Ramberg-Schmeiser (1974) (RS) GλD is an extension of Tukey's Lambda distribution (Hastings, Mosteller, Tukey & Windsor 1947). It is defined by its inverse distribution function:

$$F^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2} \qquad (1)$$

Steve Su is affiliated with School of Mathematics and Statistics at University of Western Australia and Covance Pty Ltd, Sydney, Australia. Email him at: allegro.su@gmail.com.

In (1), $0 \le u \le 1$, $\lambda_2 \neq 0$ and λ_1 ,λ_2, λ_3, λ_4 are respectively the location, inverse scale and shape parameters of the generalized Lambda distribution GλD(λ_1 ,λ_2, λ_3, λ_4). Karian, Dudewicz and MacDonald (1996) noted that GλD is defined if and only if:

$$\frac{\lambda_2}{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}} \ge 0 \text{ for } u \in [0,1]. \tag{2}$$

Another distribution known as FKML GλD also exists (Freimer, Kollia, Mudholkar, & Lin, 1988). The FKML GλD can be written as:

$$F^{-1}(u) = \lambda_1 + \frac{\frac{u^{\lambda_3} - 1}{\lambda_3} - \frac{(1-u)^{\lambda_4} - 1}{\lambda_4}}{\lambda_2} \tag{3}$$

Under (2), $0 \le u \le 1$ and λ_1 ,λ_2, λ_3, λ_4 are consistent with the interpretations in RS GλD, namely λ_1 ,λ_2 are the location and inverse scale parameters and λ_3, λ_4 are the shape parameters.

The fundamental motivation for the development of FKML GλD is that the distribution is proper over all λ_3 and λ_4 (Freimer Mudholkar, Kollia & Lin, 1988). The only restriction on FKML GλD is that $\lambda_2 > 0$.

The most commonly used technique in mixture distributional fitting is maximum likelihood estimation. This is usually achieved by using the EM algorithm for explicitly defined probability functions such as the Normal, Gamma and Exponential. In the case of implicitly defined distributions such as the GλDs, it is possible to use two ways of estimating the parameters of the mixtures, the maximum likelihood estimation using the EM algorithm and the partitioned maximum likelihood method which utilizes the complete data log likelihood. Both methods are discussed below.

GλDs Fitting Mixture Algorithm

The fitting of mixture of two GλDs is completed using the following algorithm:

Step 1

Divide the data into two parts. This can be done using a variety of clustering methods. Practical experience has shown that clustering methods such as Clara and Fanny described in Kaufman and Rousseeuw (1990) worked well in a wide range of situations. However, any clustering method that gives a reasonable classification can be used. This step provides a starting value for p in the mixture distribution equation $pf_1+(1-p)f_2$, which will be optimized later. The Clara clustering method appears to work well for a wide variety of empirical data and all fitting results in this article uses this clustering method.

To maximize the partition log likelihood this is all that is required. In the case of maximizing the log likelihood using EM algorithm, each partition of the data set additionally contains the maximum and minimum values of the entire data set as well as 1% (it is often worthwhile to explore different percentages to obtain better initial values for the maximum likelihood fitting scheme) of randomly selected data from the other group.

For example, if data sets 1 and 2 both have 100 observations, data set 1 will contain 102 observations, including 1 observation randomly selected from data set 2 and 1 maximum value from data set 2 (if it was not selected already), assuming data set 1 already contains the minimum value of the original data set. This is to ensure that the partitioned data span the entire range of the data; a necessary step because the goal is to maximize the log likelihood for the mixture data

Step 2

For each part of the data, fit a statistical distribution using maximum likelihood estimation (Su 2007a, Su 2007b).

Step 3

After the distribution fits for both parts of the data are obtained, the final parameters are estimated by maximizing the appropriate formula in (4) (for partition maximum likelihood) or (5) (for the EM algorithm approach). The initial value of p comes from step 1 and the initial values for this stage of the optimization are from step 2. The maximization is conducted numerically via the Nelder-Mead Simplex algorithm and only solutions that span the entire original data set are accepted. The formulae required in this maximization step are discussed below.

Let X, Z be the complete data, with X~ $f_1(x,\theta)$ if z = 0 and X ~ $f_2(x,\theta)$ if z=1, Then, the complete data log likelihood is given by:

$$l_c(\theta, p) = \sum_{i=1}^{n}(1-z)\begin{Bmatrix}\log(f_1(x_i,\boldsymbol{\theta_1}))+\\ \log(p)\end{Bmatrix}+z\begin{Bmatrix}\log(f_2(x_i,\boldsymbol{\theta_2}))+\\ \log(1-p)\end{Bmatrix} \tag{4}$$

Using standard statistical calculations, the conditional expectation of $l_c(\theta, p)$ given x is:

$$\sum_{i=1}^{n}T_i\begin{Bmatrix}\log(f_1(x_i,\boldsymbol{\theta_1}))\\ +\log(p)\end{Bmatrix}+S_i\begin{Bmatrix}\log(f_2(x_i,\boldsymbol{\theta_2}))\\ +\log(1-p)\end{Bmatrix} \tag{5}$$

and

$$S_i = \frac{f_2(x_i,\theta_2)(1-p)}{f_2(x_i,\theta_2)(1-p)+f_1(x_i,\theta_1)(p)} \tag{6}$$
$$1-S_i = T_i$$

where f_1 and f_2 are GλD distributions fitted to each partition of the data set and θ_1 and θ_2 representing the parameters associated with these distributions respectively. In the case of two RS GλDs mixture fits, for example, equation (4) becomes:

$$\left(\sum_{i=1}^{n_1}\log(p)+\log\left[\frac{\lambda_2}{\lambda_3 u_i^{\lambda_3-1}+\lambda_4(1-u_i)^{\lambda_4-1}}\right]\right)$$
$$+\left(\sum_{j=1}^{n_2}\log(1-p)+\log\left[\frac{\delta_2}{\delta_3 v_i^{\delta_3-1}+\delta_4(1-v_i)^{\delta_4-1}}\right]\right),$$

with $n_1 + n_2 = n$. Here the n_1 and n_2 are the number of observations in each partition of the data set and the δ_k for k = 1, 2, 3, 4 represents the parameters of the second GλD fit, similarly u_i and v_i represents the quantiles for each partition of the data set for the i^{th} observation.

All other combinations of different RS and FKML GλD fits for complete data log likelihood and maximum likelihood via EM algorithm can be found by substituting the required GλD into (4) or (5) and hence are not detailed herein.

Step 4

The parameters obtained in step 3 are then used to maximize (7). The results of this optimization process are the final parameters for the GλD mixture fits. This step was omitted in Su (2007a) but subsequent updates to the GLDEX package in R, by default, has added this optimization step for both partition and full maximum likelihood methods.

$$\sum_{i=1}^{n}\log(p(f_1(x_i,\boldsymbol{\theta_1}))+(1\text{-}p)(f_2(x_i,\boldsymbol{\theta_2}))) \tag{7}$$

Step 5

The final fitting result can be examined by plotting the result on the histogram with the fitted line, quantile plots as well as testing the goodness of fit using the Kolmogorov-Smirnov (KS) test. A two sample KS test is carried out by sampling 90% of the empirical data from the actual distributions and this is compared to equal number of data from the corresponding fitted distributions. This is repeated 1,000 times with the result of this test being the number of times the p-value exceeds 0.05 (or at a specified significance level) over 1,000 times. This will give the user an independent measure as to the adequacy of fits beyond a visual comparison.

Although this study is focused on fitting two mixtures of GλD, fitting three or more mixtures of GλD is a straightforward extension. In the case of three mixtures, it is possible to divide the data into three partitions, apply maximum likelihood estimation to each partition to find the initial values and maximize the following partition maximum likelihood or EM maximum likelihood formulae to find the parameters of the mixture distribution. To achieve this, let X, Z again be the complete data and X~$f_j(\mathbf{x},\boldsymbol{\theta})$ if z_j = 1, with j = 0, 1, 2. The proportion of the data in f_j are represented by p_j. The complete data likelihood or partition maximum likelihood is given in (8) and the conditional expectation of complete data log likelihood given **x** is given in (9).

$$
\begin{aligned}
l_c(\boldsymbol{\theta},\mathbf{p}) = \sum_{i=1}^{n} & z_0\{\log(f_0(x_i,\boldsymbol{\theta_0}))+\log(p_0)\} \\
& +z_1\{\log(f_1(x_i,\boldsymbol{\theta_1}))+\log(p_1)\} \\
& +z_2\{\log(f_2(x_i,\boldsymbol{\theta_2}))+\log(p_2)\}
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
& \sum_{i=1}^{n}\frac{f_0(x_i,\boldsymbol{\theta_0})(p_0)}{w_i}\{\log(f_0(x_i,\boldsymbol{\theta_0}))+\log(p_0)\} \\
& +\frac{f_1(x_i,\boldsymbol{\theta_1})(p_1)}{w_i}\log(f_1(x_i,\boldsymbol{\theta_1}))+\log(p_1)\} \\
& +\frac{f_2(x_i,\boldsymbol{\theta_2})(p_2)}{w_i}\log(f_2(x_i,\boldsymbol{\theta_2}))+\log(p_2)\} \\
& w_i = f_0(x_i,\boldsymbol{\theta_0})(p_0)+f_1(x_i,\boldsymbol{\theta_1})(p_1)+f_2(x_i,\boldsymbol{\theta_2})(p_2)
\end{aligned} \tag{9}
$$

Based on the parameters obtained in maximizing (8) or (9), the last step of the optimization is to maximize (10), this gives the final parameters of the mixture distribution fit.

$$
\begin{aligned}
& \sum_{i=1}^{n}\log(p_0(f_0(x_i,\boldsymbol{\theta_0}))+p_1(f_1(x_i,\boldsymbol{\theta_1}))+ \\
& p_2(f_2(x_i,\boldsymbol{\theta_2})))
\end{aligned} \tag{10}
$$

The development of partition maximum likelihood method and maximum likelihood via EM algorithm is intended to cover two different types of modeling situations. The first situation is when two distributions are distinct and disjoint, in which partition maximum likelihood would be the method of choice. The second situation is where two distributions overlap with each other in which the full maximum likelihood would be more preferable. However, this does not preclude the use of either methods in any given situation and the choice of one method over the other could still be based on more objective measures such as KS test and QQ plots.

The method presented here and in Su (2007a, Su 2007b) optimizes the maximum likelihood directly rather than use the usual method of differentiation. This is a much more efficient and reliable method of achieving the maximum likelihood rather than differentiating and solving a system of linear equations because in many cases, GλD may be undefined for certain parameter values, rendering the technique of differentiation useless. Hence, it is usually preferable to use a general purpose optimization scheme such as the Nelder-Simplex algorithm to fit GλDs.

Results

The effectiveness of using the algorithm described earlier to fit mixture of two and three generalized lambda distributions to a range of simulated and empirical data are now illustrated. The graphical displays of resulting fits are shown in Figures 1 and 2, and the numerical goodness of fit assessments are shown in Tables 1 and 2. Partition maximum likelihood method and maximum likelihood method using the EM algorithm are abbreviated as PML and ML in the outputs respectively.

In Figure 1, data set 1 is generated by 70% of Normal (mean = 10, standard deviation = 3) and 30% of exponential distributions. Data set 4 is generated by 50% of double exponential and 50% of Normal (mean = 5, standard deviation = 2) distributions. Both data sets 1 and 4 consist of 1,000 observations. Data sets 2, 3 and 5 are various data collected from the internet by the author and consist of 72, 244 and 272 observations, respectively. The data illustrated in Figure 2 is a relatively well known galaxy of white dwarf stars and consists of 7,140 observations. Numerical summaries of these data are provided in Tables 1 and 2.

The QQ plots in Figure 1 indicates that the algorithm using either partition or full maximum likelihood are convincing fits to the empirical data, this is supported by the high values indicated by the KS tests and in many cases, the theoretical moments of the fitted GλDs are quite close to the empirical data. In particular, Figure 1b demonstrates the type of distributional fits expected from using partition maximum likelihood methods; there is a tendency for the method to make a sharper split between the two data. This is reinforced in the comparison between Figure 1d and 1e, where a more abrupt separation of the two data sets can be observed in 1e using the partition maximum likelihood method. It is, however, not always true that the partition maximum likelihood will result in a jagged distributional shape; as Figure 1f shows, the resulting fit is smooth.

Overall, both methods of fitting mixtures provide a good fit to a range of data and it is recommended to examine both methods in most cases. For example, it may be preferable (due to closer match of to the moments of data and better KS test results) to use partition maximum likelihood with user defined setting for data in Figure 2, but the maximum likelihood using EM algorithm is preferred for data set 4. Clearly, no one fitting method will work the best in every case, so the choice of different methods is important to allow users to cope with different data with different tools. Sensitivity analysis using different distributional fits may also be carried out, to examine the robustness of a particular strategy under different representations of a probability distribution.

In many situations, the default setting of the GLDEX package works well. However, as known in mixture distribution modeling, the choice of initial values can have a large impact on the resulting fits. This is clearly demonstrated in Figure 2, where the default separation of the data into three parts using Clara classification scheme failed to give a very convincing fits as indicated in Figure 2a and 2b. The use of a user defined clustering regime in identifying the sub distributions (data < 100, data between 100 to 300, data > 300) leads to superior fits as shown in Figure 2c and 2d and the partition maximum likelihood with user defined data split is remarkably close to the first four moments of the empirical data.

Conclusion

This article demonstrates an algorithm to fit mixtures using the GλD distribution family. An important advantage of using GλD distribution is the elimination of the type of distributions that need to be used to model multi modal data. A critical improvement needed for all fitting methods of GλD is the search of suitable initial values. Although a fairly robust approach is provided here and in Su (2010b, 2007a, 2007b), it may be possible to directly find a set of good initial values from empirical data to speed up the optimization process and to increase the prospect of reaching a global maximum.

References

Böhning, D., & Seidel, W. (2003). Recent developments in mixture models. *Computational Statistics and Data Analysis*, *41*, 349-357.

Freimer, M., Kollia, G., Mudholkar, G., & Lin C. (1988). A study of the generalized Tukey lambda family. *Communications in Statistics- Theory and Methods*, *17*(*10*), 3547-3567.

Hastings, J. C., Mosteller, F., Tukey, J., & Windsor C. (1947). Low moments for small samples: A comparative study of order statistics. *The Annals of Statistics*, *18*, 413-426.

Karian, Z., & Dudewicz, E. (2000). *Fitting statistical distributions: The generalized lambda distribution and generalized bootstrap methods*. New York, NY: Chapman and Hall.

Karian, Z., Dudewicz, E., & McDonald, P. (1996). The extended generalized lambda distribution systems for fitting distributions to data: History, completion of theory, tables, applications, the final word on moment fits. *Communications in Statistics-Computation and Simulation*, *25*(*3*), 611-642.

Karlis, D., & Xekalaki E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, *41*, 577-590.

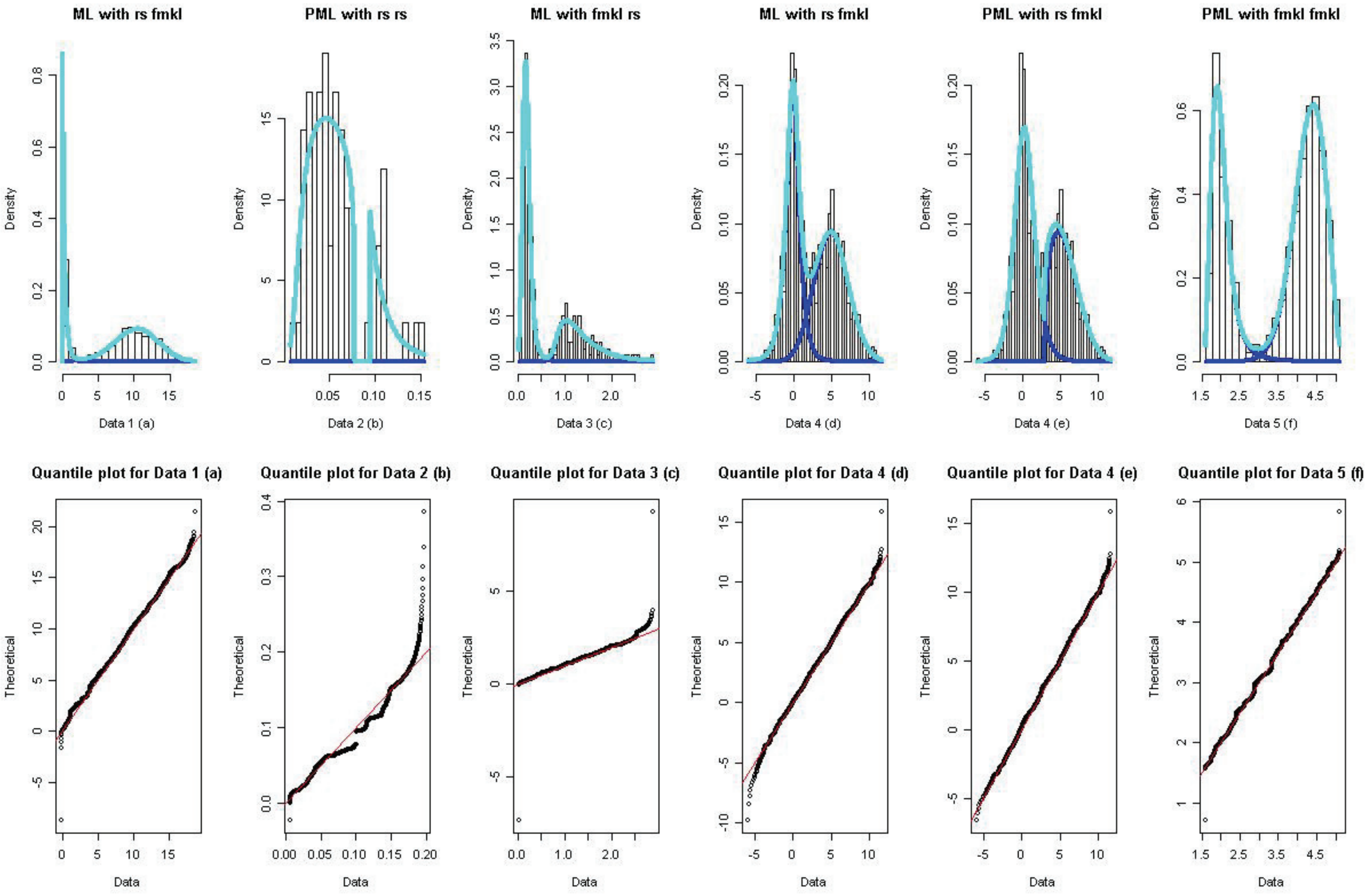
Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York, NY: Wiley.

Miloslavsky, M., & van der Laan, M. (2003). Fitting of mixtures with unspecified number of components using cross validation distance estimate. *Computational Statistics and Data Analysis*, *41*, 413-428.

Okur M., (1988). On fitting the generalised Lambda distribution to air pollution data. *Atmospheric Environment*, *22*, 2569-2572.

Ramberg, J., & Schmeiser, B. (1974). An approximate method for generating asymmetric random variables. *Communications of the Association for Computing Machinery*, *17*, 78-82.

Figure 1: Examples of Fitting Bimodal Data with a Mixture of Two Generalized Lambda Distributions



Su, S. (2010a). Fitting GLD to data via quantile matching method. In *Handbook of distribution fitting methods with R*, Z. Karian, & E. Dudewicz, Eds., 1171-1205. Boca Raton: CRC Press/Taylor & Francis.

Su, S. (2010b). Fitting gld to data using the GLDEX 1.0.4 in R. In *Handbook of distribution fitting methods with R*, Z. Karian, & E. Dudewicz, Eds., 585-608. Boca Raton: CRC Press/Taylor & Francis.

Su, S. (2005). A discretized approach to flexibly fit generalized lambda distributions to data. *Journal of Modern Applied Statistical Methods*, *4*, 408-424.

Su, S. (2007a). Fitting single and mixture of generalized lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R. *Journal of Statistical Software*, *21*(*9*), 1-17.

Su, S. (2007b). Numerical maximum log likelihood estimation for generalized lambda distributions. *Computational Statistics and Data Analysis*, *51*(*8*), 3983-3998.

Table 1: Numerical Results Indicating Goodness of Fit In Terms of First Four Moments and Resample KS Tests for Figure 1

	Data 1	(a)	Data 2	(b)	Data 3	(c)	Data 4	(d)	(e)	Data 5	(f)
Mean	7.23	7.28	0.06	0.06	0.62	0.63	2.56	2.56	2.62	3.49	3.49
Variance	26.89	26.76	0.00	0.00	0.39	0.39	10.01	10.07	9.87	1.30	1.30
Skewness	-0.17	-0.20	1.09	1.76	1.19	1.21	0.36	0.33	0.28	-0.42	-0.41
Kurtosis	1.70	1.69	3.77	12.87	3.60	3.89	2.24	2.30	2054.78	1.50	2.11
Number of times KS test p value > 0.05 out of 1,000		912		949		948		985	833		943

Figure 2: Examples of Fitting Trimodal Data with a Mixture of Three Generalized Lambda Distributions (This example illustrates how splitting data manually can improve the fit beyond the default settings.)

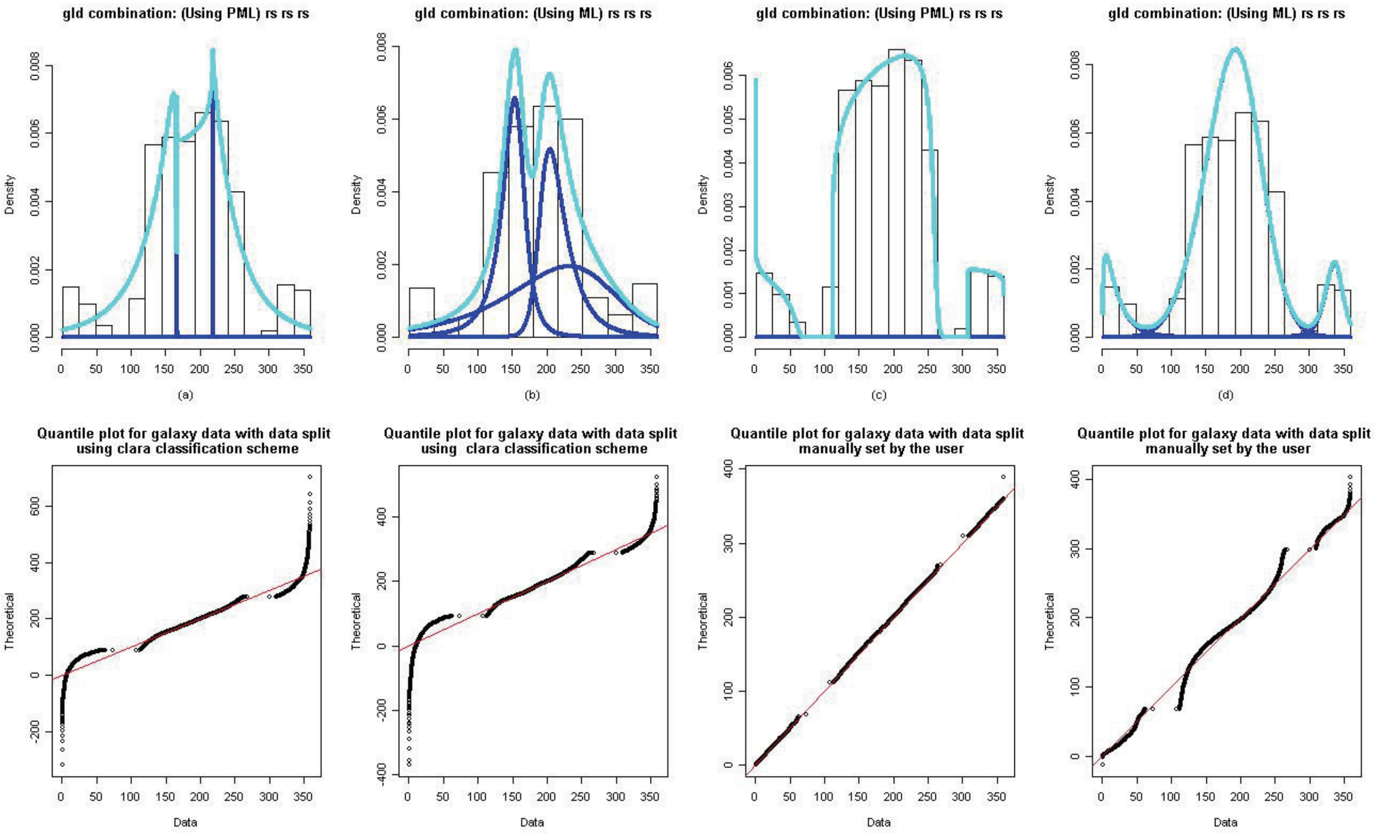


Table 2: Numerical Results Indicating Goodness of Fit In Terms of First Four Moments and Resample KS Tests for Figure 2

	Data	PML Using Clara Scheme	ML Using Clara Scheme	PML with Manual Setting	ML with Manual Setting
Mean	187.78	187.82	188.06	188.32	187.69
Variance	4870.03	5110.28	5665.51	4868.24	4946.95
Skewness	-0.18	-0.09	-4.02	-0.20	2.29
Kurtosis	3.85	7.32	NA	3.87	-1112094.77
Number of times KS test p value > 0.05 out of 1,000		850	769	938	317