CHAPTER IV
ESTIMATION AND FITTING

4.1 Introduction

In the previous chapter we have introduced the modified lambda family with the objective of considering it as a plausible model of income distribution. Supplementing the theoretical justifications given earlier for using MLF as an income model due to its versatility it is essential to establish its empirical validity by showing that members of the family fits income data. Towards this endeavour in the present chapter our attempt is to devise procedures for estimating the parameters of the MLF, establish some theoretical results that supports the use of such estimators and finally show by goodness of fit procedures that the distribution describes the data adequately. Since our aim is to substantiate the relevance of MLF as an income model, a deeper analysis of the proposed estimation procedure vis-à-vis other competing methods is not attempted. However, a short review of the existing procedures for estimating the parameters of the MLF has been conducted in Section 2, for the sake of completion. The new estimation procedure based on comparing selected characteristics in the population and in the sample is presented in Section 3. In Section 4, MLF is fitted to a real income data along with the assessment of the goodness of fit through Chisquare criterion. Finally in Section 5, the accuracy of the estimates arrived at by the proposed procedure is compared with those of the method of moments and percentiles through a simulation study.

4.2 Review of estimation techniques

Due to the mathematical form of the quantile function and the extent of the parameter space induced by the four parameters, the likely correlation between the estimates and time consuming operational problems renders the question of obtaining appropriate parameters for the MLF often a challenging task. There are many general methods of estimations prescribed
for the Ramberg and Schmeiser (1974) generalized lambda distribution mentioned in equation (2.38), but the formulas therein do not apply themselves to MLF. Hence our discussion of the present section confine only to those specifically confined to the MLF.

King and Mac Gillivray (1999) proposed the starship method which consists in
(a) transforming the data on $X$ to $F(X)$
(b) calculating the value of the Kolmogorov distance or Anderson-Darling distance for the values of $F(x)$ and the uniform distribution over $(0,1)$
(c) choosing $\lambda$-values that minimizes the distance. In a discussion of the method, the authors point out that it is of ‘numerically intensive nature’ requiring computer power to fit the data and analytical results are not available for the expected value and standard errors of the estimates.

Tarsitano (2005) considered the quantile function

$$Q(p) = \lambda_1 + \lambda_2 p^{\lambda_3} - \lambda_4 (1 - p)^{\lambda_5}$$

which contains five parameters, $\lambda_1$ for location, $\lambda_2$ and $\lambda_3$ representing scale and $\lambda_4$ and $\lambda_5$ describing the shape. The model according to the author contains MLF and therefore his general conclusions about the estimates remain valid for the latter. Various methods of estimation discussed are percentile method that matches five selected sample percentiles with corresponding theoretical percentiles, method of moments, matching probability weighted moments $E[Q(p_i)]$, $i = 0,1,2,3,4$ with the sample counterparts

$$t_0 = \frac{1}{n} \sum_{j=1}^{n} C_j, \quad t_i = \sum_{j=i+1}^{n} C_j \prod_{r=1}^{j-i} (j-r) / \prod_{r=0}^{i} (n-r), \quad i = 1,2,3,4$$

where $C'$s are midpoints of class intervals, minimum distance method using Cramer – Von Mises statistic, that minimizes
$D = \sum (p_i - F(X_i))^2$

where $F$ is the estimated p value that would generate the observation $X_i$, 
maximum likelihood estimates obtained by minimizing the negative log-
likelihood

$L = -\sum_{i=1}^{k} \eta_i \left( F(X_i) - F(X_{i-1}) \right)$

and the Pseudo least square approach based on

$X_i = E(X_i) + \varepsilon_i$

The simulation study for comparing the different methods for 36 configurations revealed that the minimum distance and probability weighted moments approaches gave the 'worst' results. Further, the method of maximum likelihood had given results 'slightly better than these obtained by minimum distance, but not by an amount of any practical importance besides both being computationally demanding. The percentile, moment and Pseudo least squares were reported to give desirable results.

King and Mac Gillivray (2006) introduced the notion of spread functions

$S_F(p) = F^{-1}(p) - F^{-1}(1 - p)$

in defining shape functionals

$r(p) = \frac{F^{-1}(p) + F^{-1}(1 - p) - 2m}{S(p)}$

and

$\eta(p, q) = F^{-1}(p) + F^{-1}(1 - p) - \frac{[F^{-1}(q) + F^{-1}(1 - q)]}{S(q)}$, $\frac{1}{2} < q < p < 1$

by which estimates that minimize the distance between sample and population values of the functionals were proposed. For the MLF short tails were found to be problematic in the estimation procedure. Due to the various limitations pointed out in starship method, maximum likelihood, minimum distance and shape functionals, consideration will be given in the present study to computationally simple and reasonably accurate methods using
percentiles and moments. First we present a new procedure for estimation by matching selected characteristics in the population and in the sample.

4.3 New Estimation Procedure

The method proposed in the present section resembles that of the classical methods of selected points, with the difference that the points chosen here is derived by matching the basic characteristics of the distribution viz. location, dispersion, skewness and kurtosis with those in the sample. The choice of the characteristics ensures that the parameter values determined there from corresponds to the true values that provide the same location, scale and shapes with a reasonable degree of accuracy. The accuracy results empirically from the criterion for optimization and theoretically from the asymptotic properties established in the sequel. The measures of location, dispersion, skewness and kurtosis involved in the new estimation procedure are the quantile based measures viz. Median, Quartile deviation, Galton's coefficient of skewness and Moor's kurtosis measure respectively. The expressions relating to these measures were obtained in equations (3.25), (3.27), (3.28) and (3.29) in the previous chapter. Hence the method of estimation of the parameters \( \lambda_i, i = 1, 2, 3, 4 \) in the model (3.22) is by solving for the \( \lambda_i \)'s from the equations obtained by setting (3.25), (3.27), (3.28) and (3.29) respectively equal to the corresponding measures in the sample. To accomplish this, we define the \( pth \) quantile corresponding to a random sample \( (X_1, X_2, \ldots, X_n) \) of observations on \( X \) as the \( pth \) quantile \( \hat{\xi}_p \) of the sample distribution function \( F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \) where \( I(.) \) is the indicator function. Denoting the sample median, quartile deviation, measures of skewness and kurtosis by \( m, r, s \) and \( t \) based on \( \hat{\xi}_p \) the problem is to solve the simultaneous equations

\[
\lambda_1 + \frac{1}{\lambda_2} \left[ \frac{0.5^h - 1}{\lambda_3} - \frac{0.5^h - 1}{\lambda_4} \right] = m
\]
\[ \frac{1}{2\lambda_2} \left[ \frac{0.75^{b} - 0.25^{b}}{\lambda_2} + \frac{0.75^{b} - 0.25^{b}}{\lambda_4} \right] = r \] 
\[ \lambda_4 \left[ 0.75^{b} - 2 \times 0.5^{b} + 0.25^{b} \right] - \lambda_3 \left[ 0.75^{b} - 2 \times 0.5^{b} + 0.25^{b} \right] = s \] 
\[ \lambda_4 \left[ 0.75^{b} - 0.25^{b} \right] + \lambda_3 \left[ 0.75^{b} - 0.25^{b} \right] = t \]

It may be noted that the above equations are non-linear and therefore, end
up with more than one quadruple of \((\lambda_1, \lambda_2, \lambda_3, \lambda_4)\) values that satisfy them.

The solutions being unrestricted, some set of solutions may not be within the
range prescribed for the \(\lambda\)'s in Chapter 3, so that a proper probability
distribution will not result. Secondly when more than one set arise as solution
with admissible range, there is the question of some criterion that
distinguishes the best solution. A solution to the first problem is to discard \(\lambda\)
values that do not fall within the parameter space. Answering how a choice
be made when multiple admissible solutions occur can be made with the aid
of an optimality criterion. One simple way of devising such a criterion is to
ensure that the difference between the estimated values and the sample
values of the measures of location, dispersion, skewness and kurtosis are
within a preassigned small value. Since there are four such differences the
criterion is prescribed as
\[ e = \max \left( |M - m|, |R - r|, |S - s|, |T - t| \right) < \varepsilon \]
for some positive \(\varepsilon\), sufficiently small.

The computation of the solutions are carried using the command
'FindRoot' in Mathematica, which requires a set of initial values for the \(\lambda\)'s to
initiate the solution. While different initial values in some cases, may give
different solutions, the \(\varepsilon\) criterion is invoked to find the best among them.
Thus empirically the method proposed leads to a unique solution within the
parameter space that nearly reproduces population characteristics that matches these found in the sample.

Our next step is to show that the procedure can also be justified from a theoretical stand point. Consider a sample of size $n$ from a one-dimensional distribution of the continuous type with distribution function $F(x)$ and density function $f(x)$. Let $\xi_p$ denote the $p$th quantile, $0 < p < 1$ and suppose that in some neighbourhood of $\xi_p$, $f(x)$ is continuous and has a continuous derivative $f'(x)$. Then it is known that (Serfling, R. J. (1980)) the $p$th sample quantile $z_p$ is asymptotically normal $\left( \xi_p, \frac{1}{f'(\xi_p)} \sqrt{\frac{pq}{n}} \right)$. Further as a special case the median of the sample $m$ is a strongly consistent estimator of the population median $M = \xi_{\frac{1}{2}}$ and

$$m = \xi_{\frac{1}{2}}$$

is asymptotically normal $N\left( M, \frac{1}{4} n \left[ f''(M) \right]^2 \right)$. Then $m$ belongs to the class of CAN estimators. In the same manner,

$$r = \frac{1}{2} (z_{0.75} - z_{0.25})$$

is also a CAN estimator with distribution

$$N\left( R, \frac{1}{64 n} \left( \frac{3}{f^2(Q(0.75))} - \frac{2}{f(Q(0.25)) f(Q(0.75))} + \frac{3}{f^2(Q(0.25))} \right) \right)$$

for large values of $n$. The results in Serfling (1980) concerning the functions of quantiles can be adapted suitably to the result that $(s, t)$ is consistent for $(S, T)$ and $\sqrt{n} (s - S, t - T)^*$ has asymptotic bivariate normal distribution with mean $(0, 0)^*$ and dispersion matrix $\phi'(c) A [\phi'(c)]^*$ where

$$A = (\sigma_{ij}), \quad \sigma_{ij} = \frac{i(i - j)}{64 f(E_i) f(E_j)}, \quad i \leq j \quad \phi(c) = (S, T)^*$$

and $^*$ denotes the transpose. The expected values of $m - M$, $r - R$, $s - S$ and $t - T$ being zero for large samples, our estimating equations $m = M$, $r = R$, $s = S$ and $t = T$ provide values that agrees with the above
expected values with small variations implied by the consistency of the estimators.

4.4 Fitting MLF to income data

Having set the background material for inference, the next important stage in model building is to test the model against the observations for adequacy. For the purpose, we consider the income data from Arnold (1983); referred to as 'Texas counties data' consisting of 157 observations. Each observation represents the total personal income accruing to the population of one of the 254 counties in Texas in 1969. The 157 included in the present data set represent all the Texas counties in which total personal income exceeds $20,000,000.

From the observations, the sample characteristics required for our estimation procedure are computed as

\[ m = 46.3, \quad r = 37.25, \quad s = 0.5651, \quad t = 2.4362 \]

Substitution in equations (4.1) through (4.4) and following the \( \varepsilon \) criteria in the computational process gave the following admissible solutions (using Mathematica)

\[ \hat{\lambda}_1 = 27.3207, \quad \hat{\lambda}_2 = 0.0441223, \quad \hat{\lambda}_3 = 3.86057, \quad \hat{\lambda}_4 = -1.19399 \]

Of the various possible values of the \( \lambda \)'s for different initial values above estimates are optimum according the \( \varepsilon \) criteria that gave the maximum error of \( \varepsilon = 0.0282 \). In order to assess the appropriateness of the MLF with the above parameter values for the given observations, a frequency distribution was formed by classifying the data into 10 intervals (the class interval was taken unequal for larger incomes to accommodate a reasonable frequency) and the corresponding expected frequencies using the above estimates of the parameters are exhibited in Table 4.1.
### Table 4.1

**Lambda distribution fit for Texas Counties Data**

<table>
<thead>
<tr>
<th>Class intervals</th>
<th>Observed frequency</th>
<th>Expected frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;30</td>
<td>39</td>
<td>41.8895</td>
</tr>
<tr>
<td>30-40</td>
<td>27</td>
<td>25.9491</td>
</tr>
<tr>
<td>40-50</td>
<td>13</td>
<td>15.7135</td>
</tr>
<tr>
<td>50-60</td>
<td>13</td>
<td>10.6738</td>
</tr>
<tr>
<td>60-70</td>
<td>6</td>
<td>7.78014</td>
</tr>
<tr>
<td>70-80</td>
<td>11</td>
<td>5.94857</td>
</tr>
<tr>
<td>80-100</td>
<td>8</td>
<td>8.53515</td>
</tr>
<tr>
<td>100-200</td>
<td>15</td>
<td>18.1278</td>
</tr>
<tr>
<td>200-500</td>
<td>16</td>
<td>12.121</td>
</tr>
<tr>
<td>&gt;500</td>
<td>9</td>
<td>10.2614</td>
</tr>
</tbody>
</table>
The chi-square goodness of fit test provides $\chi^2 = 7.884$ that do not reject the hypothesis that observations follow the MLF. The histogram of the data and the frequency curve from the expected frequencies for the various class intervals are shown in Fig 4.1. Thus it is clear that MLF could be used as a model of incomes and that our method of estimation provides estimates of reasonable accuracy.

**Figure 4.1**

Histogram with MLF fit

Though the new procedure of estimation has both empirical and theoretical support, it is of interest to know how it fares in comparison with some of the standard methods. For reasons noted in Section 2 of the current chapter where review of the different approaches were taken up, for comparison purposes we choose the method of percentiles and method of moments.
Since our expressions for the sample statistics are non-linear in the parameters analytic derivations of the standard errors of the \( \lambda \)'s are difficult to obtain. Hence a quick assessment of the sampling variations in the estimates for the given data is not possible. Therefore we have conducted a simulation study to assess the standard errors of the estimates for comparison with other methods. These are presented in the next Section.

4.5 Comparison with the Methods of Moments and Percentiles

The most popular approach for estimating parameters in various forms of the lambda distribution is based on matching the first four moments of the empirical data with those of the population. In our approach, measures \((M, R, S, T)\) provide alternatives to the moment induced quantities \((\mu, \sigma^2, \beta_1, \beta_2)\). We have preferred the former set in view of the following.

(i) The method of moments is restricted to distributions possessing fairly light tails because they must have finite moments. Heavy tails usually observed in empirical distribution does not support such a premise. Moreover, the sample moments are sensitive to extreme observations or other contaminants in the data and sampling variability in higher moments can be large. In income data usually the interval lengths are taken unequal then the estimated moments cannot be subjected to corrections for grouping, resulting in highly biased estimates.

(ii) Existence of moments requires restrictions on the parameter space that are not always satisfied by the solution. This is not a necessary condition for using \(M, R, S\) and \(T\).

(iii) The quantities \(M, R, S\) and \(T\) can be found graphically and further has the advantage of being operative without the necessity of knowing every measurement.

(iv) \(S\) and \(T\) are invariant under location and scale and \(R\) is location invariant.
(v) The numerical values $(S, T)$ show the same pattern of behaviour as $(\beta_1, \beta_2)$, except for the difference in size of the numerical values. Thus from a theoretical standpoint the method proposed in the present study has several advantages over the method of moments.

The percentile method to fit MLF to a given data consists in equating four suitably sample quantiles to their MLF counterparts and solving the resulting equations for $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$. The four sample statistics are defined by

\[
\hat{\lambda}_1 = \hat{Q}(0.5)
\]
\[
\hat{\lambda}_2 = \hat{Q}(0.9) - \hat{Q}(0.1)
\]
\[
\hat{\lambda}_3 = \frac{\hat{Q}(0.5) - \hat{Q}(0.1)}{\hat{Q}(0.9) - \hat{Q}(0.5)}
\]
\[
\hat{\lambda}_4 = \frac{\hat{Q}(0.75) - \hat{Q}(0.25)}{\hat{Q}(0.9) - \hat{Q}(0.1)}
\]

These sample statistics have the following interpretations. $\hat{\lambda}_1$ is the sample median; $\hat{\lambda}_2$ is the inter-decile range, i.e., the range between the 10th percentile and 90th percentile; $\hat{\lambda}_3$ is the left-right tail-weight ratio, a measure of relative tail weights of the left tail to the right tail (distance from median to the 10th percentile in the numerator and distance from 90th percentile to the median in the denominator); and $\hat{\lambda}_4$ is the tail weight factor or the ratio of the inter-quartile range to the inter-decile range, which is $\leq 1$ and measures how great tail weight is (values close to 1.00 indicate the distribution is not spread out greatly in its tails, while values close to 0 indicate the distribution has long tails).

For MLF, these measures are obtained as
\[ \rho_1 = \lambda_1 + \frac{1}{\lambda_2} \left[ \frac{0.5^{\lambda_1} - 1}{\lambda_3} - \frac{0.5^{\lambda_4} - 1}{\lambda_4} \right] \]

\[ \rho_2 = \frac{1}{\lambda_2} \left[ \frac{0.9^{\lambda_1} - 0.1^{\lambda_1}}{\lambda_3} + \frac{0.9^{\lambda_4} - 0.1^{\lambda_4}}{\lambda_4} \right] \]

\[ \rho_3 = \frac{\lambda_4 \left[ 0.5^{\lambda_1} - 0.1^{\lambda_1} \right] + \lambda_3 \left[ 0.9^{\lambda_1} - 0.5^{\lambda_1} \right]}{\lambda_4 \left[ 0.9^{\lambda_1} - 0.1^{\lambda_1} \right] + \lambda_3 \left[ 0.5^{\lambda_1} - 0.1^{\lambda_1} \right]} \]

\[ \rho_4 = \frac{\lambda_4 \left[ 0.75^{\lambda_1} - 0.25^{\lambda_1} \right] + \lambda_3 \left[ 0.75^{\lambda_1} - 0.25^{\lambda_1} \right]}{\lambda_4 \left[ 0.9^{\lambda_1} - 0.1^{\lambda_1} \right] + \lambda_3 \left[ 0.9^{\lambda_1} - 0.1^{\lambda_1} \right]} \]

Now solving the equations \( \rho_1 = \hat{\rho}_1, \rho_2 = \hat{\rho}_2, \rho_3 = \hat{\rho}_3 \) and \( \rho_4 = \hat{\rho}_4 \) using mathematica we get the estimates \( \hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3 \) and \( \hat{\lambda}_4 \). The computational aspects discussed for the new method are valid for the percentile method as the four equations are nonlinear.

To assess the performance of the above three competing methods we have conducted a simulation study by generating samples of size 33 (to accommodate the quantiles) from MLF with parameters \( \lambda_1 = 13.7, \lambda_2 = 0.2, \lambda_3 = 0.4 \) and \( \lambda_4 = 0.01 \). The parameters were then estimated using the three methods. The same procedure was repeated for samples of size 66. The bias and S.E are presented in Table 4.2.

Based on the simulation studies carried over several samples have revealed the following features associated with the various methods.

The absolute bias in the new method tend to decrease with increasing sample size for the estimates of all the four parameters, though the reduction in bias is not considerable. The estimates are thus more or less stable. This is confirmed by almost the same type of behaviour seen in the case of the mean square error as well.
Table 4.2:
Bias and Mean Square Error

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>Parameters</th>
<th>Sample Size</th>
<th>Absolute Bias</th>
<th>Mean Square Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New Method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>$n = 33$</td>
<td>0.0407</td>
<td>0.163</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.03942</td>
<td>0.0927</td>
<td></td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$n = 33$</td>
<td>0.00088</td>
<td>0.000697</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.00084</td>
<td>0.000682</td>
<td></td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>$n = 33$</td>
<td>0.0608</td>
<td>0.0116</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.0564</td>
<td>0.01365</td>
<td></td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>$n = 33$</td>
<td>0.0612</td>
<td>0.0346</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.0563</td>
<td>0.0322</td>
<td></td>
</tr>
<tr>
<td><strong>Method of Percentiles</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>$n = 33$</td>
<td>0.0459</td>
<td>0.1522</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.0622</td>
<td>0.1543</td>
<td></td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$n = 33$</td>
<td>0.0007</td>
<td>0.0008</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.0099</td>
<td>0.00377</td>
<td></td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>$n = 33$</td>
<td>0.0571</td>
<td>0.0169</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.01499</td>
<td>0.0184</td>
<td></td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>$n = 33$</td>
<td>0.0622</td>
<td>0.0397</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.0165</td>
<td>0.03796</td>
<td></td>
</tr>
<tr>
<td><strong>Method of Moments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>$n = 33$</td>
<td>0.23695</td>
<td>0.1894</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.1854</td>
<td>0.1451</td>
<td></td>
</tr>
<tr>
<td>$\lambda_2$</td>
<td>$n = 33$</td>
<td>0.0314</td>
<td>0.0063</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.0296</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>$\lambda_3$</td>
<td>$n = 33$</td>
<td>0.2422</td>
<td>0.1459</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.2074</td>
<td>0.1088</td>
<td></td>
</tr>
<tr>
<td>$\lambda_4$</td>
<td>$n = 33$</td>
<td>0.1234</td>
<td>0.062</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n = 66$</td>
<td>0.1186</td>
<td>0.0597</td>
<td></td>
</tr>
</tbody>
</table>
The percentile method shows larger bias for $\lambda_1$, fluctuating bias for $\lambda_2$ and decreasing bias for $\lambda_3$ and $\lambda_4$. Generally the numerical value of the bias is seen larger than that of the new method. The mean square errors are lower for the new method, showing that there is more concentration of the estimates about the location measure.

Of the three methods, the method of moments fares the worst having produced considerably larger bias and mean square errors than the other two methods. Over all the impression gathered from the simulation study is that our method compares favourably and at times better than the method of percentiles and moments.