Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 3983 – 3998
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Numerical maximum log likelihood estimation for generalized lambda distributions

Steve Su*

The George Institute for International Health, Level 24, 207 Kent Street, Veritas Building, Sydney, New South Wales 2000, Australia

Received 19 September 2005; received in revised form 7 June 2006; accepted 7 June 2006
Available online 5 July 2006

Abstract

This paper presents a two-step procedure using the method of moment or percentile to find initial values and then maximize the numerical log likelihood to fit the appropriate generalized lambda distribution to data. This paper demonstrates the use of this procedure to fit well-known statistical distributions as well as some empirical data. Overall, the use of numerical maximum log likelihood estimation is a valuable alternative among existing methods of fitting. It provides not only convincing results in terms of quantile plots and goodness of fit tests but also has the advantage of a lower variability in its parameter estimation compared to the existing starship (King and MacGillivray, 1999) and method of moment (Karian and Dudewicz, 2000) fitting schemes.

Keywords: Fitting distributions; Prior distributions; Empirical data analysis

1. Introduction

An essential problem in data analysis is to find a probability distribution that will adequately fit the empirical data. Considerable literature exists in this area, ranging from the parametric work of generalized lambda distribution (GλD) (Freimer et al., 1988; Karian and Dudewicz, 2000; King and MacGillivray, 1999; Lakhany and Massuer, 2000; Okur, 1988; Ozturk and Dale, 1985; Ramberg and Schmeiser, 1974; Ramberg et al., 1979; Su, 2005a) to non-parametric work of kernel density estimation (Silverman, 1985; Wegman, 1972). Surprisingly, no current work exists on using numerical maximum log likelihood method to fit GλD to data, and this paper will demonstrate the use of this technique.

The paper begins with a literature review on the existing methods of GλD parameters estimation, which progressively result in the development of this new method. Results of the application of the new methods on real life data are then presented and the paper concludes with a discussion on the shortcomings of this new method.

2. Literature review

This literature review begins with the basic theory of GλD and discusses some of the fitting methods reported in literature. The literature review then presents two methods that appear to give promising results. These two methods are extended and discussed in the method section.

* Tel.: +61 (0) 421 840 586; fax: +61 (02) 9657 0301.
E-mail addresses: allegro.su@gmail.com, ssu@george.org.au (S. Su).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.008

2.1. Basic background

The Ramberg–Schmeiser (1974) (RS) GλD is an extension of Tukey's lambda distribution (Hastings et al., 1947). It is defined by its inverse distribution function:

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

In Expression (1), $0 \leqslant u \leqslant 1$, $\lambda_2 \neq 0$ and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are, respectively, the location, inverse scale, skewness and kurtosis parameters of GλD$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. In particular, Karian et al. (1996) noted that GλD is defined if and only if:

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

Another distribution known as FMKL GλD also exists, due to the work of Freimer et al. (1988). The FMKL GλD can be written as

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

Under Expression (3), $0 \leqslant u \leqslant 1$ and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ 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 FMKL GλD is that the distribution is proper over all λ_3 and λ_4 (Freimer et al., 1988). This adds convenience to users who wish to program this function as there are fewer restrictions on the values of λ_3 and λ_4. The only restriction on FMKL GλD is $\lambda_2 > 0$.

In practice, it is possible to use either the FMKL or RS GλD. Due to the wide range of shapes GλD possesses, for example: U shaped, bell shaped, triangular, and exponentially shaped distributions and its simplicity, it has been used in Monte Carlo simulations (Hogben, 1963), the modeling of empirical distributions (Okur, 1988; Ramberg et al., 1979), in response time modeling (Au-Yeung et al., 2004: Kumaran and Achary, 1996), supply chain (Ganesalinggam and Kumar, 2001) and in the sensitivity analysis of robust statistical methods (Shapiro et al., 1968).

Other research works on GλD concentrate on estimating the parameters of the GλD from empirical data. Perhaps the most common approach has been to use method of moments as demonstrated in Ramberg et al. (1979) and Karian and Dudewicz (2000), Karian et al. (1996). These works covered only the RS GλD and the principal aim is to find a GλD with parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ that matches closely with the first four moments of the empirical data. Karian and Dudewicz (2000) also discussed an alternative method which matches the parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ of the RS GλD with the first four percentiles of the data set. This is a variation on the same theme of the moment matching method but one in which Karian and Dudewicz (2003, 2000) reported can produces better fits than the methods of moment under the RS GλD.

In a different line of work, Ozturk and Dale (1985) used a version of least squares estimation to find the parameters of RS GλD. They derived the squared distance between empirical data points with the expected values of the order statistics, and numerically minimized this measure using Nelder–Simplex method to derive parameter estimates for the RS GλD.

The literature recognizes that matching the first four moments or using the "least squares" method by Ozturk and Dale (1985) does not necessarily produce a good fit to the data (Karian and Dudewicz, 2000; Lakhany and Massuer, 2000). In the case of the moment matching method, different parameters of the GλD can have similar first four moments. In the case of the "least squares" method by Ozturk and Dale (1985), the goal of minimizing the squared distance between empirical data points with the expected values of the order statistics of GλD does not necessarily coincide with the formal goodness of fit objective such as the Kolmogorov–Smirnov goodness-of-fit test.

It is precisely the need to assess the resulting fit with the goodness of fit objective that King and MacGillivray (1999) used the starship methods. In the starship method, grid points comprising of $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ aimed at covering a wide range of GλD, were calculated from the sample quantiles. Then, for each of the grid points the theoretical GλD was transformed into uniform distribution to find the goodness of fit statistics such as Anderson–Darling test statistics or Kolmogorov–Smirnov (KS) test statistics. The set of grid points with the lowest Anderson–Darling statistics was then chosen as the initial values for optimization, usually through the Nelder–Simplex algorithm. The resulting values from the optimization scheme are the parameter estimates of the GλD, given by starship method.

Lakhany and Massuer (2000) suggested a variation of using re-sampling method combined with the method of moments and the goodness of fit test via the FMKL GλD. They first generated initial values for the method of moment matching via quasi random number generator (for example, the Sobol sequence generator, Bratley and Fox, 1988), and then found the set of values $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ that matched optimally (through the Nelder–Simplex algorithm) with the first four moments from the data. This set of values was then evaluated through a goodness of test statistic such as adjusted KS test statistics demonstrated in their work. Under this method, any solution that results in a p-value $>$ 0.05 is accepted. Lakhany and Massuer (2000) commented that this method is much more efficient time-wise than the starship method developed by King and MacGillivray (1999) and allows for automatic restarts from different initial values to help to find a distribution that will adequately fit the data. The use of p-values in the optimization scheme, however, can be somewhat problematic. The deficiency of p-values is well known, since failure to reject does not mean the hypothesis is true since it may be that the sample size is too small to be able to detect differences between the empirical and fitted data. Conversely, rejection of the hypothesis does not mean the fitted model is inappropriate, as the user may have a different purpose to fitting the data other than to satisfy the goodness of fit criteria.

The current literature appears to endorse both the method of percentiles under RS GλD and the method of moments under the FMKL GλD which offer great flexibility of fit. These distributions have also been used previously by Su (2005a) to flexibly fit a wide range of empirical distributions successfully and this paper will offer an alternative, more definitive fit using the numerical maximum log likelihood method.

A detailed discussion of the method of percentiles using the RS GλD and the method of moments using FMKL GλD is outlined below. These discussions are necessary as they form the first step of finding initial values for maximizing the numerical log likelihood.

2.2. *Method of percentiles using the RS GλD*

The following is obtained directly from Karian and Dudewicz (2000). For a given data set X with values $x_1, x_2, \ldots, x_n$, the pth percentile defined by Karian and Dudewicz (2000) is $\hat{\pi}_p = y_r + k\,(y_{r+1} + y_r)$, where $Y = y_1, y_2, \ldots, y_n$ are sorted values of X in ascending order and r is the truncated value of $(n+1) \times p$ with k being $(n+1) \times p - r$.

Instead of using the first four moments, the following statistics are used:

$$\begin{aligned}
\hat{\rho}_1 &= \hat{\pi}_{0.5},\\
\hat{\rho}_2 &= \hat{\pi}_{1-v} - \hat{\pi}_v,\\
\hat{\rho}_3 &= \frac{\hat{\pi}_{0.5} - \hat{\pi}_v}{\hat{\pi}_{1-v} - \hat{\pi}_{0.5}},\\
\hat{\rho}_4 &= \frac{\hat{\pi}_{0.75} - \hat{\pi}_{0.25}}{\hat{\rho}_2},
\end{aligned} \tag{4}$$

where v is an arbitrary number from 0 to 0.25 and it is taken as 0.1 in this paper to be consistent with Karian and Dudewicz (2000).

The relationship between the theoretical $\rho_1, \rho_2, \rho_3, \rho_4$ and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ in the RS GλD is as follows

$$\begin{aligned}
\rho_1 &= F^{-1}(0.5) = \lambda_1 + \frac{0.5^{\lambda_3} - 0.5^{\lambda_4}}{\lambda_2},\\
\rho_2 &= F^{-1}(1-v) - F^{-1}(v) = \frac{(1-v)^{\lambda_3} - v^{\lambda_4} + (1-v)^{\lambda_4} - v^{\lambda_3}}{\lambda_2},\\
\rho_3 &= \frac{F^{-1}(0.5) - F^{-1}(v)}{F^{-1}(1-v) - F^{-1}(0.5)} = \frac{(1-v)^{\lambda_4} - v^{\lambda_3} + (0.5)^{\lambda_3} - (0.5)^{\lambda_4}}{(1-v)^{\lambda_3} - v^{\lambda_4} + (0.5)^{\lambda_4} - (0.5)^{\lambda_3}},\\
\rho_4 &= \frac{F^{-1}(0.75) - F^{-1}(0.5)}{\rho_2} = \frac{(0.75)^{\lambda_3} - (0.25)^{\lambda_4} + (0.75)^{\lambda_4} - (0.25)^{\lambda_3}}{\rho_2}.
\end{aligned} \tag{5}$$

The condition $-\infty < \rho_1 < \infty$, $\rho_2 \geqslant 0$, $\rho_3 \geqslant 0$, $\rho_4 \in [0, 1]$ must also be true, which is a direct consequence of the definition of $\rho_1, \rho_2, \rho_3, \rho_4$. In Karian and Dudewicz (2000), a fit for the GλD is found by solving Expression (6) through the use of tables. This can also be solved numerically via Newton–Raphson method

$$\left|\hat{\rho}_3 - \rho_3\right| \leqq 10^{-6}, \quad \left|\hat{\rho}_4 - \rho_4\right| \leqq 10^{-6}. \tag{6}$$

In the extended method described below, however, the following minimization scheme in Expression (7) is used. Once λ_3, λ_4 are obtained, λ_1, λ_2 can be obtained directly via Expression (5).

$$\sqrt{\left(\hat{\rho}_3 - \rho_3\right)^2 + \left(\hat{\rho}_4 - \rho_4\right)^2}. \tag{7}$$

2.3. *Method of moments under the FMKL GλD*

In an alternative approach, Lakhany and Massuer (2000) used the method of moments for the FMKL GλD. The following are extracts from Lakhany and Massuer (2000):

For a given data set X with values $x_1, x_2, \ldots, x_n$, the ith moment α_i is defined in Expression (8)

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{\sum_{i=1}^{n} x_i}{n}, \\
\hat{\alpha}_2 &= \frac{\sum_{i=1}^{n} \left(x_i - \hat{\alpha}_1\right)^2}{n}, \\
\hat{\alpha}_3 &= \frac{\sum_{i=1}^{n} \left(x_i - \hat{\alpha}_1\right)^3}{n\left(\hat{\alpha}_2\right)^{1.5}}, \\
\hat{\alpha}_4 &= \frac{\sum_{i=1}^{n} \left(x_i - \hat{\alpha}_1\right)^4}{n\left(\hat{\alpha}_2\right)^2}.
\end{aligned} \tag{8}$$

Putting $a = 1/\lambda_2$ and $b = \lambda_1 - 1/\lambda_1\lambda_2 + 1/\lambda_2\lambda_4$, with $Y = (X - b)/a$, using $E\left(X^k\right) = \int_0^1 \left(F^{-1}(u)\right)^k \mathrm{d}u$ and binomial expansion gives Expression (9)

$$\begin{aligned}
s_k &= E\left(Y^k\right), \\
s_k &= \int_0^1 \left(\frac{u^{\lambda_3}}{\lambda_3} - \frac{(1-u)^{\lambda_4}}{\lambda_4}\right) \mathrm{d}u, \\
s_k &= \int_0^1 \sum_{j=0}^{k} \binom{k}{j} (-1)^j \left(\frac{u^{\lambda_3(k-j)}}{\lambda_3^{k-j}} - \frac{(1-u)^{\lambda_4 j}}{\lambda_4^j}\right) \mathrm{d}u, \\
s_k &= \sum_{i=0}^{k} \binom{k}{j} \frac{(-1)^j}{\lambda_3^{k-j} \lambda_4^j} \beta\left(\lambda_3(k-j) + 1, \lambda_4 j + 1\right).
\end{aligned} \tag{9}$$

In Expression (9), $\beta(*)$ denotes beta function. Note that both arguments of the beta function must be positive, implying that $\min(\lambda_3, \lambda_4) > -1/k$ if the distribution is to have finite kth moments. The kth central moment (except for the first which is the mean) of the distribution $F^{-1}(u)$ denoted as μ_k are hence given in Expression (10)

$$\begin{aligned}
\mu_1 &= \frac{1}{\lambda_2}(s_1) - \frac{1}{\lambda_2\lambda_3} + \frac{1}{\lambda_2\lambda_4}, \\
\mu_2 &= \frac{1}{\lambda_2}\left(s_2 - s_1^2\right), \\
\mu_3 &= \frac{1}{\lambda_2^3}\left(s_3 - 3s_1s_2 + 2s_1^3\right), \\
\mu_4 &= \frac{1}{\lambda_2^4}\left(s_4 - 4s_1s_3 + 6s_1^2s_2 - 3s_1^4\right).
\end{aligned} \tag{10}$$

The theoretical α_3 and α_4 are given in Expression (11)

$$\alpha_3 = \frac{s_3 - 3s_1s_2 + 2s_1^3}{(s_2 - s_1)^{3/2}},$$
$$\alpha_4 = \frac{s_4 - 4s_1s_3 + 6s_1^2s_2 - 3s_1^4}{(s_2 - s_1)^2}. \tag{11}$$

The same method now follows as from Lakhany and Massuer (2000). They propose to find λ_3, λ_4 by minimizing Expression (12), where $\hat{\alpha}_3$ and $\hat{\alpha}_4$ are sample values using sample moments

$$\sqrt{(\hat{\alpha}_3 - \alpha_3)^2 + (\hat{\alpha}_4 - \alpha_4)^2}. \tag{12}$$

Once λ_3, λ_4 is determined, λ_1, λ_2 can be found using Expression (13)

$$\lambda_2 = \frac{\sqrt{(s_2 - s_1^2)}}{\hat{\alpha}_2},$$
$$\lambda_1 = \hat{\alpha}_1 + \frac{1}{\lambda_2}\left(\frac{1}{\lambda_3 + 1} - \frac{1}{\lambda_4 + 1}\right). \tag{13}$$

3. Method

The method in this paper uses quasi random numbers and the percentile method from Karian and Dudewicz (2000) for RS GλD and the method of moments for FMKL GλD to find initial values. These initial values are then used to maximize the numerical log likelihood to find the parameters of the appropriate GλD. The algorithm is described below.

1. Specify a range of initial values for λ_3, λ_4, and the number of initial values to be selected. Here, the λ_3, λ_4 are set by default to range from -1.5 to 1.5 for the RS GλD percentile method and -0.25 to 1.5 for the FMKL GλD method of moment. These default values are from author's clinical experiences and appear to work well in most situations. It is possible to change these initial values if desired.

The quasi random number generator used is based on the work of Hong and Hickernell (http://www.mcqmc.org/Software.html) and the scrambling method of Owen (1995) and Faure and Tezuka (2000). The algorithm uses Niederreiter generator matrices in base 2. This code is available from the beta resample library in Splus 6.0 and scrambling methods are applied so that the numbers generated fills uniformly onto the λ_3, λ_4 two-dimensional space. It is not necessary to use the above quasi random number generators, other common quasi random number generators such as Halton and Sobol sequences (available in R) can also be used. By default, 10 000 of such initial values are chosen and used in step 2.

2. Evaluate λ_1, λ_2 for each of the initial values λ_3, λ_4. Remove all the sets of initial values that do not:

(a) Result in a legal parameterization of GλD or
(b) Span the entire region of the data set.

Among the sets of initial points not excluded by step 2, find a set of initial values that produce the lowest value in Expression (7) or (12), this set of initial values will then be used in the optimization process.

3. Calculate quantiles u_i from the initial values for RS and FMKL GλD. This can be done by solving Expressions (1) and (3) numerically.

4. Once u_i is obtained, substitute u_i into the following numerical log likelihood (14) and (15) depending on whether RS or FMKL GλD is used. These numerical log likelihoods can be obtained by applying the chain rule to differentiate Expressions (1) and (3) to obtain $f(x_i)$ and apply the logarithm on the joint distribution of $f(x_i)$, assuming independence of $f(x_i)$

$$\mathrm{ML_{RS}} = \sum_{i=1}^{n} \log\left[\frac{\lambda_2}{\lambda_3 u_i^{\lambda_3 - 1} + \lambda_4(1 - u_i)^{\lambda_4 - 1}}\right], \tag{14}$$

$$\mathrm{ML}_{\mathrm{FMKL}} = \sum_{i=1}^{n} \log\left[\frac{\lambda_2}{u_i^{\lambda_3-1} + (1-u_i)^{\lambda_4-1}}\right]. \tag{15}$$

5. The optimal result can be obtained via the Nelder–Mead Simplex algorithm or another suitable numerical optimization algorithm. It is always desirable to find another set of initial values in the optimization process to check the result obtained is a reasonable solution. The final fitting result can be examined by plotting: histogram with the fitted line and quantile plot as well as testing the goodness of fit using the resample KS tests.

The resulting fits obtained in this step are named as the revised percentile method of the RS GλD under maximum likelihood estimation (RPRS.ML) and the revised method of moment under the FMKL GλD under maximum likelihood estimation (RMFMKL.ML). These terms reflect the slight modifications involved in steps 1 and 2 in using the established methods described in 2.2 and 2.3 to find initial values for the maximum likelihood estimation.

4. Results

The analysis below is divided into three parts. To illustrate the performance of maximum likelihood method, the first part compares the sampling variance and bias of maximum likelihood estimation, starship and method of moment in fitting five FMKL GλDs for a range of sample sizes at 25, 50, 100, 200 and 400. The parameters of five FMKL GλDs are chosen to represent five classes of FMKL GλD as described in Freimer et al. (1988) and these are shown in Table 1.

The second part is a theoretical comparison between data fitting methods with well-known statistical distributions. To assess the quality of the GλD approximation, the average absolute difference in theoretical quantiles between the fitted GλD and the well-known statistical distributions are given. Quantile plots are also given to provide users with a visual comparison.

The third part shows the fitting method over some real life data. To assess the quality of fit in addition to quantile plots, a thousand two sample KS tests are carried out by comparing 90% of the original data (randomly selected) with randomly generated data from the same fitted GλD. The sample size can be 90%, 95% of the original data or whatever the percentage the user chose and the number of tests can be altered by user. This paper chooses 90% of the data and a resample scheme of 1000 times as a point of illustration. The number of times the p-value exceeds 0.05 is recorded and this test gives the user an independent measure as to the adequacy of fits beyond a visual comparison. This test is known as the Monte Carlo KS test in this paper. The use of this test rather than a comparison of a simple goodness of fit measure is to give a more stringent assessment of the goodness of fit to see if the sample generated from the data are close to those generated from the fitted distribution. Naturally, the use of goodness of fit as a measure for quality of fit would bias favorably towards methods that seek to maximize goodness of fit. In fact, it is a circular logic. The use of goodness of fit to assess the quality of fits used in this paper will not suffer from this problem, but it needs to be clear that the objective of fit in this paper is not to maximize the goodness of fit, and so it may not always be as high as starship method (STAR) which maximizes the goodness of fit of data transformed by the distribution function to the uniform.

4.1. *Performance of maximum likelihood estimation*

The results of the variance of the parameters and the mean of absolute bias (mean in absolute difference between the fitted parameters and actual parameters) are investigated over 500 simulations for each of the five different FMKL GλDs and at sample sizes of 25, 50, 100, 200 and 400. The results are given as bar charts in Figs. 1 and 2. Each row

Table 1
Five FMKL GλDs used to demonstrate the sampling variance and bias of maximum likelihood estimation

λ_1	λ_2	λ_3	λ_4
0	1	0.5	0.6
0	1	2	0.5
0	1	1.5	1.5
0	1	2.5	1.5
0	1	3	3

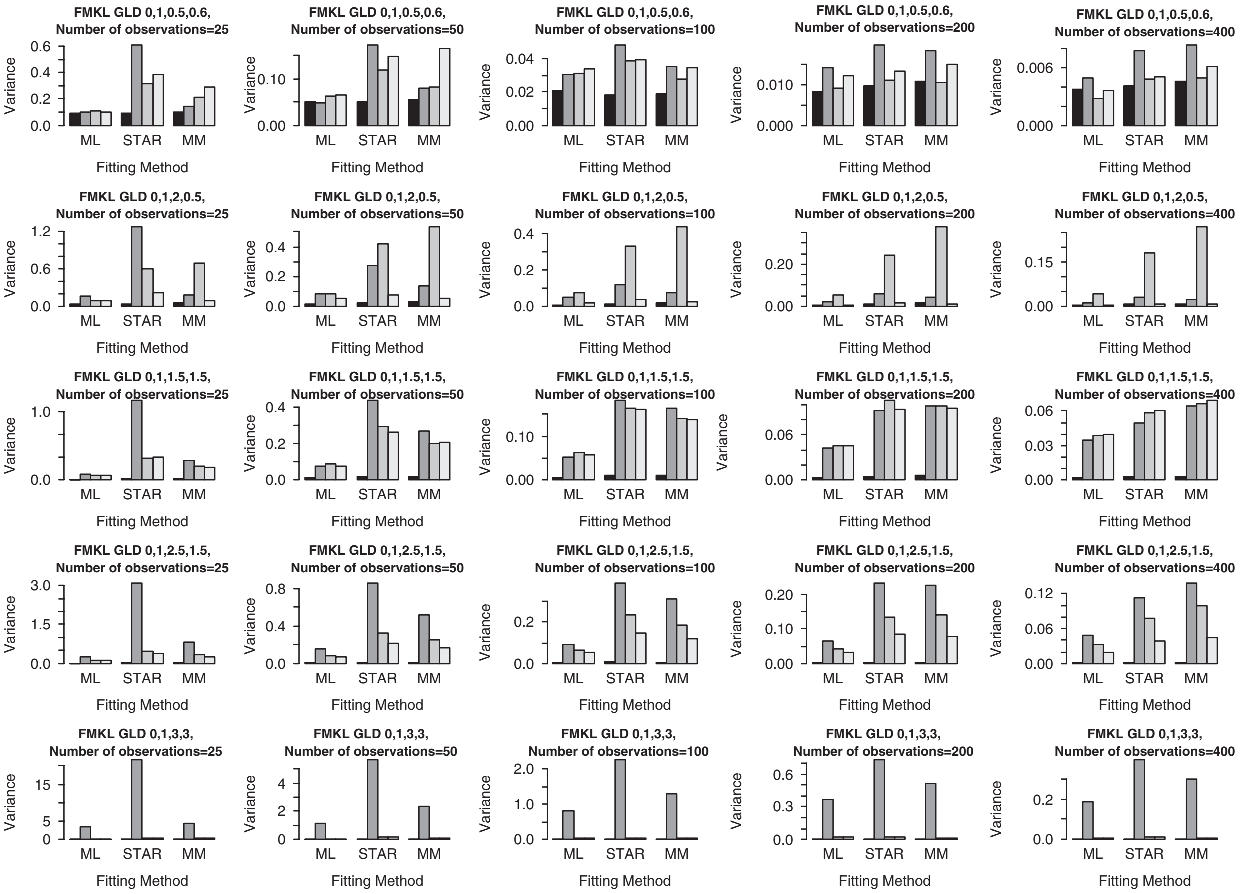

Fig. 1. This figure shows a comparison of the variance of the parameters between fitting methods over five different FMKL GλDs. The bars from left to right for each of the method group ("ML", "STAR" or "MM") represent λ_1, λ_2, λ_3, λ_4. The methods considered are maximum likelihood estimation (ML), starship method (STAR) and method of moments (MM). Overall, the maximum likelihood estimation has lower variability in its parameters estimation and this suggests the ML approach is more efficient than the other fitting methods. Each row of the above graph represents five different classes of FMKL GλD and each column represents sample size at 25, 50, 100, 200 and 400. The results reported in each individual graph above are taken from 500 fitting results over 500 simulation runs.

of the graphs in these figures represents each of the five different FMKL GλDs and each column represents sample sizes at 25, 50, 100, 200 and 400. The simulation was repeated for an additional 500 runs and there are no discernible differences in results, suggesting the size of simulation study is sufficient.

Fig. 1 shows that the variance of the parameters under maximum likelihood estimation is the lowest among the method of moment and starship methods. However, in terms of mean absolute bias, there appear to be no major differences between different methods of fitting as shown in Fig. 2.

4.2. *Comparison with theoretical distributions*

Figs. 3, 4 and Table 2 shows the resulting fits of RPRS.ML, RMFMKL.ML and STAR on well-known statistical distributions. Using the fitting method described above, RPRS.ML and RMFMKL.ML are very close to the actual distribution as shown in Fig. 3. While there are some slight differences over the peak of the curve for the approximations of $F_{6,25}$, there are very little differences towards the tails of the distributions. These good approximation behaviors of the fitted GλD are further confirmed in Fig. 4, which shows an excellent agreement of quantiles between fitted GλD and these statistical distributions. As a further check to reconcile the quantile plots, Table 2 shows the mean of absolute difference between the quantiles of the fitted GλD with the well-known statistical distributions. All the methods have

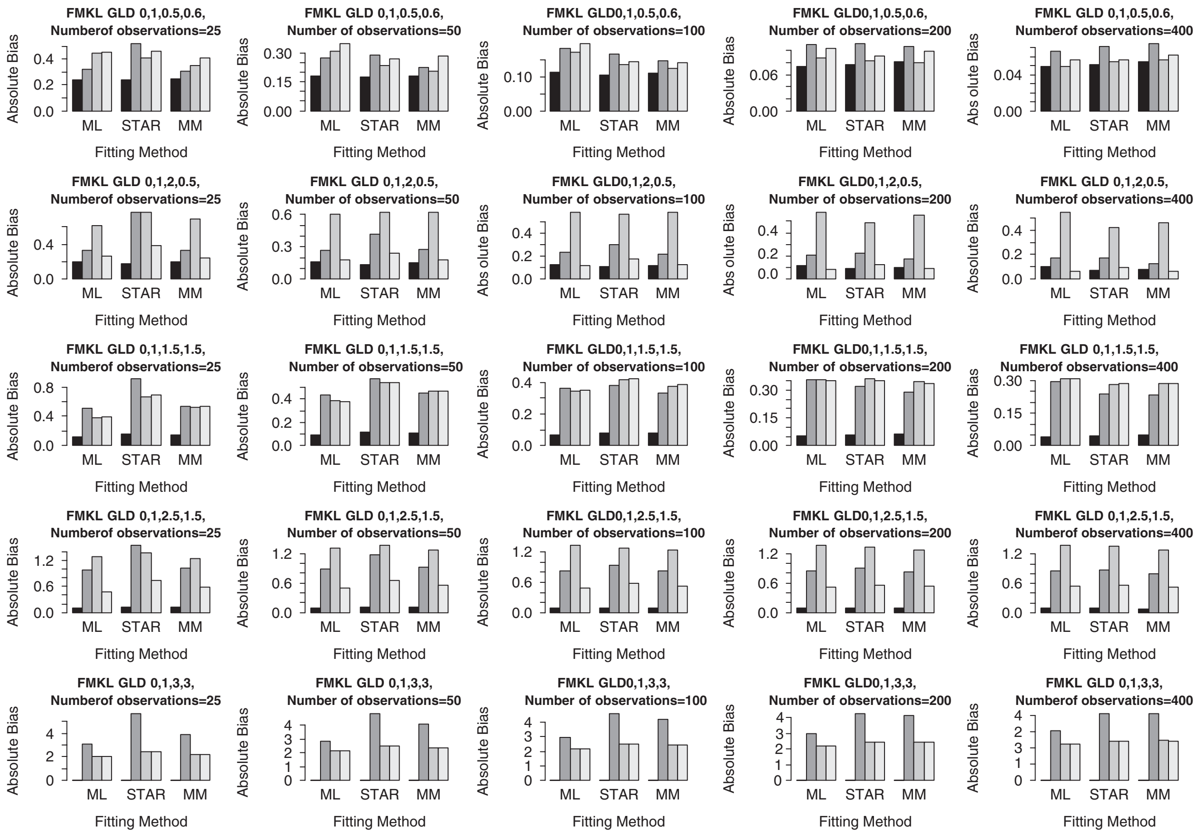

Fig. 2. This figure shows a comparison of the average absolute bias of the parameters between fitting methods over five different FMKL GλDs. The bars from left to right for each of the method group ("ML", "STAR" or "MM") represent λ_1 , λ_2, λ_3, λ_4. The methods considered are maximum likelihood estimation (ML), starship method (STAR) and method of moments (MM). There are largely no differences in the degree of average absolute bias between different methods.

quite low deviations, suggesting the approximations are very good. Table 2 also shows that there is no one method that is always superior to the other. For example, RMFMKL.ML has the lowest value for gamma(5,3) approximation but it has the highest value in beta(3,3) approximation.

The real interest of the method of this paper is not in the fitting of theoretical distributions. In the theoretical simulation, it is possible to compare between the actual and approximate distributions, but not so in practice. In practice, the data obtained are often messier and may not have a nice range of the values from the true underlying distribution; therefore, it is always necessary to test the method of fitting under a range of different real life examples to assess its fitting capabilities on empirical data.

4.3. Dataset used

The datasets used in here were supplied by research works of Sabri Hassan and Victoria Clout at School of Accountancy in Queensland University of Technology, Australia. Additional dataset on the monthly water pipe leakage repair costs during 2000–2004 is from my consulting work with a water utility firm in Queensland, Australia.

The dataset by Sabri Hassan is based on 44 Australian extractive industries firms, listed on the ASX (Australian Stock Exchange) from 1998 to 2001. The dataset used is based on the mean value of each individual company over four years. Market to Book values (sh.mtb), transparency (sh.transp), and size of the firm (sh.size) variables were extracted and used in this demonstration. There are 176 observations in this data set and the Monte Carlo KS test for 1000 runs based on a random sample of 160 observations (90%) of the data set and the fitted distribution is reported below.

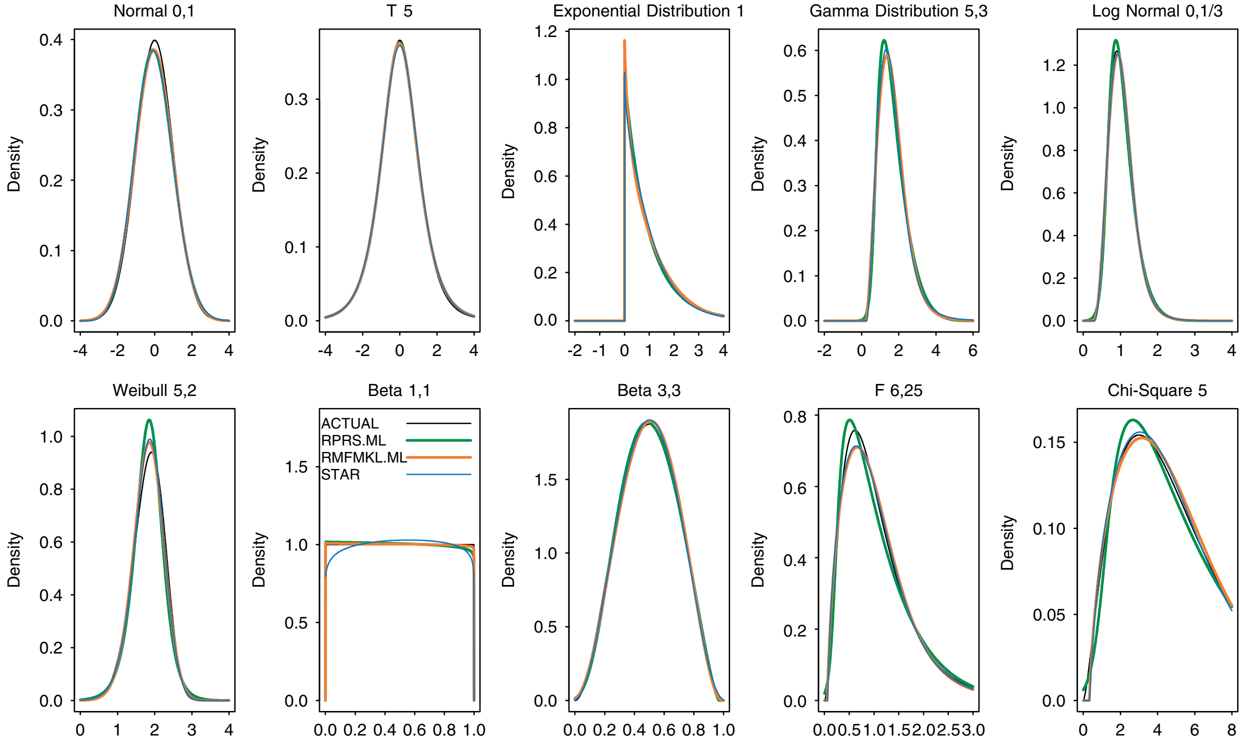

Fig. 3. Demonstrating the GλD distribution fits to well-known statistical distributions.

Victoria Clout's data consisted of 361 US firms, listed on the S&P500. The selection requirements were December year-end firms for the 1977–1995 period. Similarly, the data used is based on the mean values for each company over the 12 years period. Market to Book ratio (vc.mbr), leverage (vc.leverage), log of size (vc.size), tangible assets (vc.tang) were used in this demonstration. There are 143 observations in this data set and the Monte Carlo KS for 1000 runs based a random sample of 130 observations (90%) of the data set and the fitted distribution is given below.

The following examples are designed to demonstrate the application of the numerical maximum log likelihood method which can fit alternative, convincing distributions other than suggested by the starship method. It is also intended to give a balanced view and present situations where the objective of maximizing the goodness of fit may not be congruent to maximizing the numerical log likelihood. While the method in this paper is very close to the starship method in most cases, the following example will present two sporadic cases where this is not true.

Table 3 shows that in most cases, the numerical maximum log likelihood and starship methods give very convincing fits to the empirical data; this is supported by Figs. 5 and 6. A striking example of the difference between the numerical maximum log likelihood and starship can be seen in vc.tang and sh.transp, where different fits appear to capture different features of the data set represented by different bin-widths of the histograms. This is supported by Fig. 6, where the RMFMKL.ML appears to be out of synch with the vc.tang data at lower quantiles but fit the data better than RPRS.ML or STAR at higher quantiles. In the case of sh.transp, all the distribution fits capture the data well at different parts of the quantile but none of them appear to be an excellent fit to the data. Additionally, it is worthwhile to examine the fitting behavior of these GλDs to some extreme values as shown in sh.mtb. Only the part below zero is considered as there is only one extreme maximum value at 74.09. Fig. 7 shows that the GλD fits are still convincing at these extreme values even with only 28 data points below zero in sh.mtb. While the vc.leverage looks like a candidate for examining the extreme values, this data set only has four extreme values above 300 and 1 value below 0 so there are too little data to allow a good assessment in this case.

The example of a low goodness of fit as in the case of sh.transp can pose problems to the researcher if there is no prior information to suggest which shape is more "correct". In these situations, researcher may argue that more

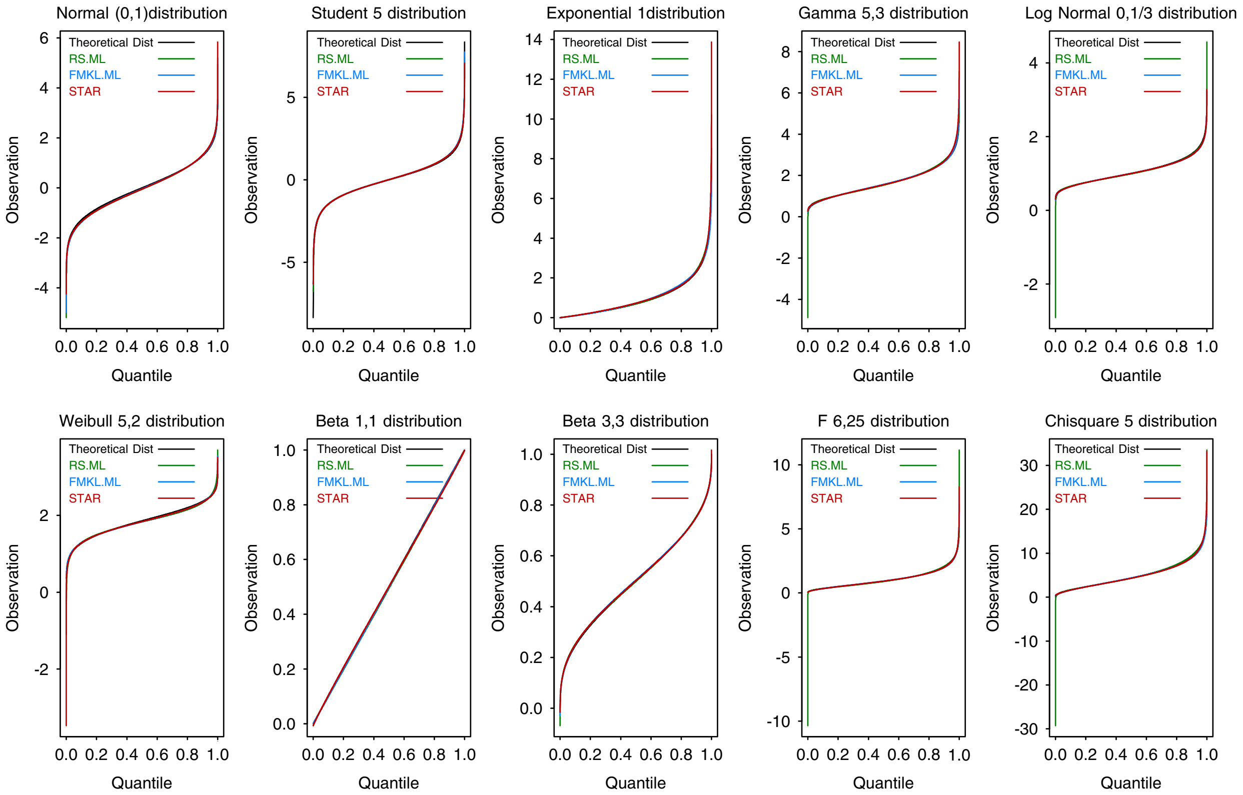

Fig. 4. Quantile plots showing the quantiles between fitted GλD and well-known statistical distributions for maximum likelihood estimation and starship method.

Table 2
Mean of absolute difference in quantiles between fitted GλD with the well-known statistical distributions

Distribution	RPRS.ML	RMFMKL.ML	STAR
normal(0,1)	0.0491	0.0441	0.0495
student(5)	0.0312	0.0353	0.0295
exp(1)	0.0316	0.0367	0.0440
gamma(5,3)	0.0256	0.0125	0.0315
lognormal(0,1/3)	0.0119	0.0125	0.0108
weibull(5,2)	0.0317	0.0177	0.0169
beta(1,1)	0.0042	0.0017	0.0042
beta(3,3)	0.0017	0.0034	0.0022
$f(6, 25)$	0.0265	0.0152	0.0118
chisq(5)	0.0835	0.0771	0.1034

Table 3
Monte Carlo KS test results testing the goodness of fit of fitted distribution on the empirical data

Data set	RPRS.ML	RMFMKL.ML	STAR
vc.leverage	907	920	933
vc.mbr	943	903	932
vc.size	918	906	917
vc.tang	903	773	911
sh.transp	0	0	270
sh.mtb	690	616	745

A value close to 1000 indicates high level of confidence of a good fit.

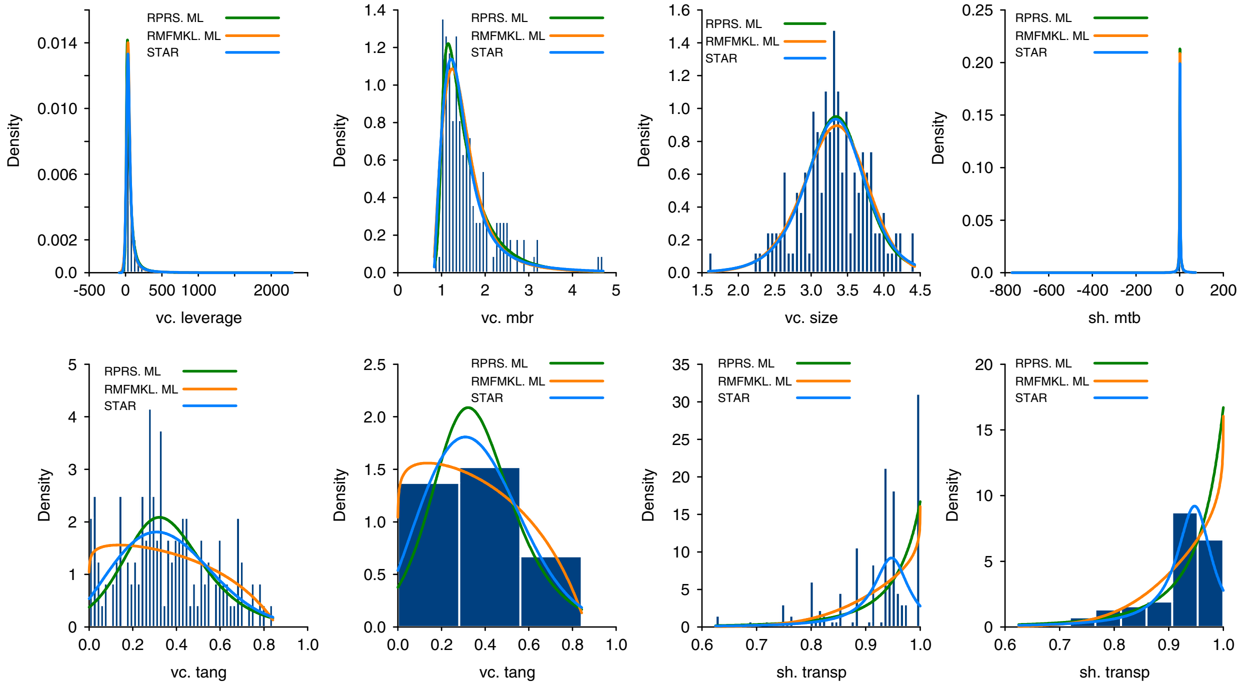

Fig. 5. Demonstrating the distribution fits of some data sets. This example intends to show in most cases, starship and numerical maximum log likelihood methods are quite similar but in some cases may capture different aspects of the data as in the case of vc.tang and sh.transp.

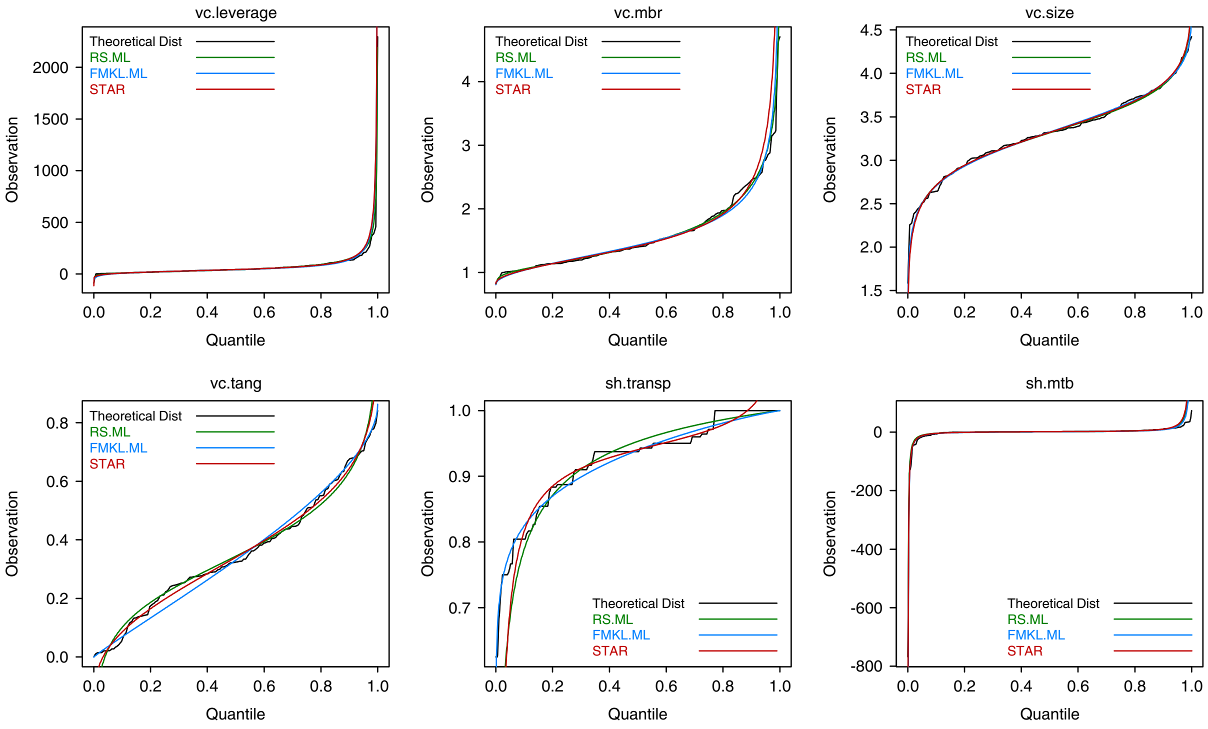

Fig. 6. This figure shows the quantile plots for six data sets. The quantiles from all the distributional fits are very close to the quantiles of the data set.

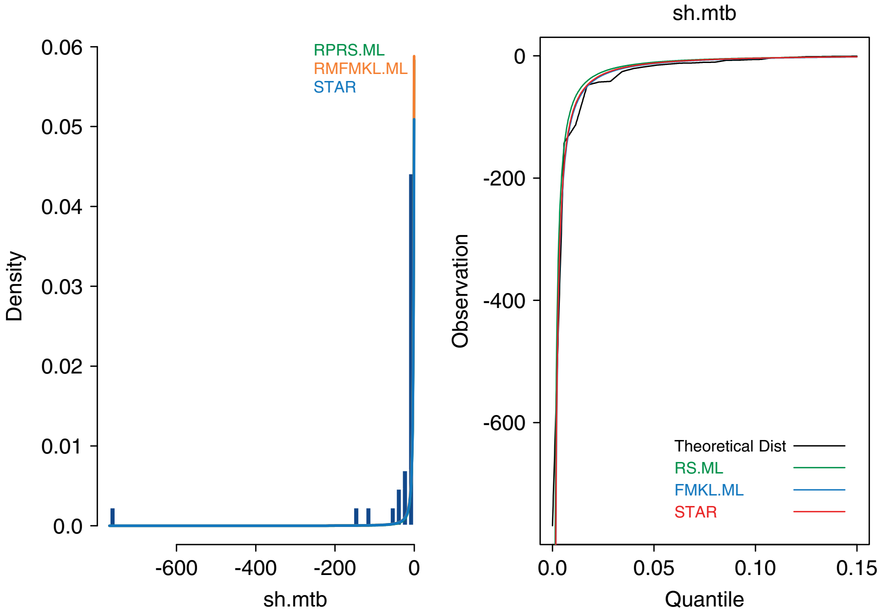

Fig. 7. This figure shows the quantile and corresponding histogram plots for the extreme values in sh.mtb. The distribution fits in these extreme values are quite convincing, especially in the light of the small amount of data in these regions.

data collection is required before a more conclusive distribution can be found. This small example also shows that the maximization of numerical log likelihood objective is not necessarily congruent to the KS goodness of fit objective and this may leave researchers with difficult choices.

As an aside, it is important that both RPRS.ML and RMFMKL.ML are considered in fitting distribution to data. In some cases, the researchers might want the fitted distribution, in addition to maximizing the numerical log likelihood, to also have the closest mean and variance and other moments to the data set and neither RPRS.ML nor RMFMKL.ML is superior to the other for this purpose. For example, the data set sh.transp has mean and variance of 0.92 and 0.0058, respectively. This is very close to RMFMKL.ML fit which result in respective theoretical mean and variance of 0.92 and 0.0048. RPRS.ML comes second with mean and variance of 0.91 and 0.027. The starship method gives a theoretical mean of 0.86 and undefined variance. In particular, RMFMKL.ML has skewness and kurtosis of −1.35 and 2.16, respectively, fairly close to the actual data result of −1.45 and 2.19. RPRS.ML does not have defined skewness and kurtosis in this case.

In the case of vc.mbr, however, the mean and variance of RPRS.ML are 1.61 and 0.43, very close to the actual mean and variance of 1.61 and 0.41. The fit by RMFMKL.ML results in mean and variance of 1.63 and 0.97, but this is still better than the STAR fit which has mean of 1.72 and undefined variance. The skewness and kurtosis of RPRS.ML fit in this case are 2.80 and 16.18 which is somewhat close to the actual data result of 2.08 and 6.01, compared to the other fits, which do not possess skewness and kurtosis statistics. These little examples show that neither RMFMKL.ML nor RPRS.ML is superior in matching the moments of the data set and they should, therefore, be used in conjunction with each other to give researchers a wider choice of the appropriate distribution for their study.

4.4. Real life example: application of modeling water pipe leakage repair costs for a water utility firm

A fundamental difficulty for many water utility firms is to model their cost distributions to allow budget setting and setting prices to recover their costs in the long run. Once the distribution is found, the probability of cost exceeding a certain amount can be found. This example is designed to illustrate the importance of using an alternative method of fitting other than starship method. The GλD fits by the monthly costs of repairs are quite convincing when plotted on the histogram for all three methods as shown in Fig. 8. The goodness of fit analysis and the method that best matches

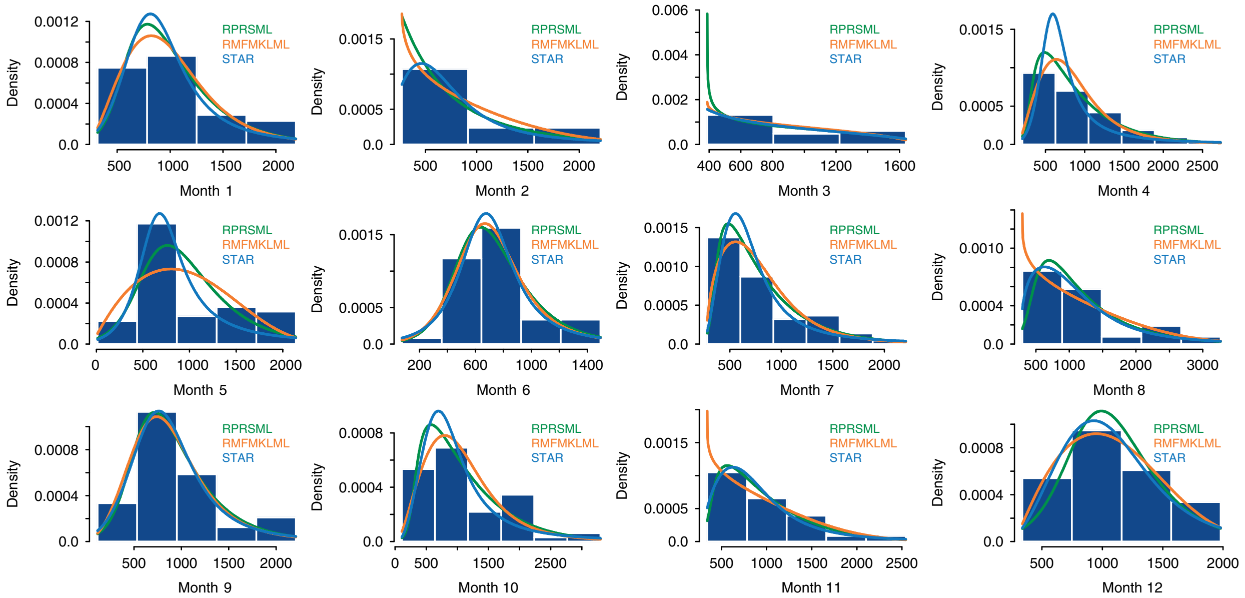

Fig. 8. Demonstrating the distribution fits of water repair data sets by month.

Table 4
Monte Carlo KS goodness of fit tests results for water repair costs over 1000 runs

Month	RPRS.ML	FMKL.ML	STAR	Closest four moments match
January	880	858	901	RPRS.ML
February	633	433	754	FMKL.ML
March	468	498	509	RPRS.ML
April	759	581	823	RPRS.ML
May	524	537	813	FMKL.ML
June	859	889	887	RPRS.ML
July	891	824	900	FMKL.ML
August	824	863	901	FMKL.ML
September	890	892	899	RPRS.ML
October	919	863	924	RPRS.ML
November	864	814	905	FMKL.ML
December	676	820	829	FMKL.ML

A value close to 1000 indicates high level of confidence of a good fit. The method that has the closest match to the mean, variance, skewness and kurtosis of data is shown on the far right column.

the four moments of the data set for each month are displayed in Table 4 and the quantile plots are exhibited in Fig. 9. The quantile plots show there is largely good agreement of fitted GλD with the starship method not capturing the data towards the higher quantiles in months 4, 5, 10 and 11. It appears that the maximum likelihood estimation is perhaps a generally better fit from these visual comparisons.

Very often, in analyzing cost distributions, such as in Statistical Activity Cost Analysis (Su, 2005b; Falta and Wolf, 2004; Willett, 1991) the distribution that best captures the moments of the data is preferred since there is often a desire to reconcile the use of these fitted distributions with traditional statistical analysis which will use the mean and variance of the data set as estimates to the underlying population. As shown in Table 4, the goodness of fits under maximum likelihood estimation are mostly very high and have closer theoretical moments to the empirical moments of the data compared to starship methods. These results are shown in Table 5.

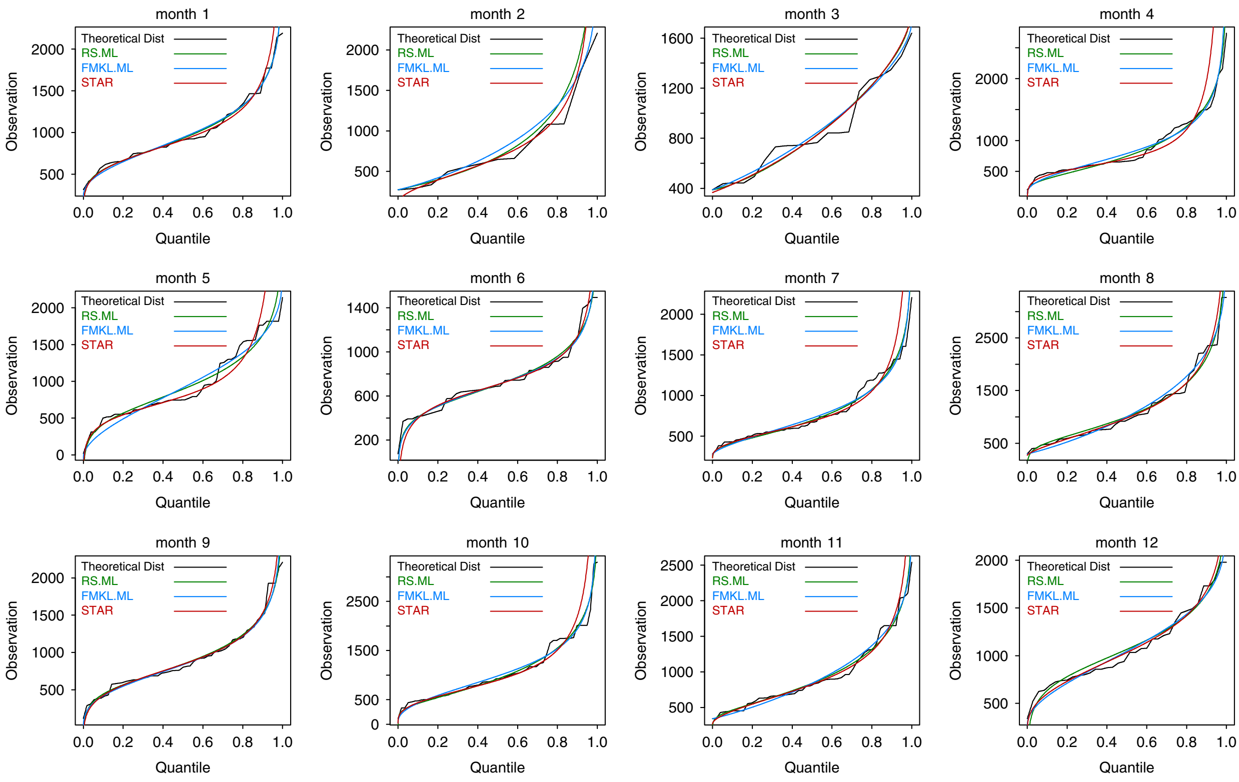

Fig. 9. This figure shows the quantile plots for monthly pipeline repair cost data. The quantiles from all the distributional fits are very close to the quantiles of the data set. In some cases, the starship method deviates from the actual quantiles, this is especially evident in month 4, 5, 10 and 11.

Table 5
Comparing the mean, variance, skewness and kurtosis of GλD fits and its corresponding monthly cost repair data

Repair data category	Data				The moments of the GλD fits that has the closest match to the data set			
	Mean	Variance	Skewness	Kurtosis	Mean	Variance	Skewness	Kurtosis
January	1013.40	194475.96	0.99	3.56	1015.95	196920.82	1.30	2046.46
February	832.18	307896.32	1.27	3.67	895.23	291573.39	1.20	4.36
March	864.31	132675.42	0.57	2.21	868.10	136440.45	0.56	2.28
April	899.76	261341.80	1.44	5.06	888.60	285466.38	1.70	−238432.28
May	949.38	252025.30	0.58	2.35	948.36	244922.13	0.33	2.51
June	745.85	95354.37	0.75	3.60	746.08	98090.89	1.10	6.34
July	813.09	160601.18	1.22	4.08	817.36	187429.54	2.58	22.24
August	1176.77	538251.06	1.24	3.93	1183.27	540301.84	1.08	3.93
September	921.64	213209.44	1.01	3.72	926.75	213994.73	1.34	6.79
October	1095.83	432779.08	1.35	4.95	1087.14	434846.88	1.40	5.57
November	959.63	255264.62	1.19	3.79	961.94	255222.46	1.04	3.78
December	1085.73	170322.06	0.60	2.51	1083.18	166416.15	0.44	2.76

5. Shortcomings of the RPRS.ML and RMFMKL.ML

All methodologies have their shortcomings, and the method devised here is no exception. The design of the RPRS.ML and RMFMKL.ML can suffer from the following deficiencies.

1. *Different results in different runs for the same settings*: RPRS.ML and RMFMKL.ML is based on re-sampling methods over the specified range of initial values, hence different runs will result in different initial values being

chosen. This is the reason sampling is based on scrambled quasi random sampling (Hong and Hickernell, 2002; Owen, 1995) available from the Splus beta resample library, so that the values span evenly throughout the ranges each time. In most cases, there are no dramatic changes between each run; however, situations do occur when the one run results in a slightly better fit than other runs. This is a problem directly related to GλD, as two GλDs with very similar shapes can often have quite different parameters.

2. *Optimization method converges falsely or do not converge*: This is a problem associated with all numerical optimization schemes, rather than related to this method directly. The program written for RPRS.ML and RMFMKL.ML allows for the quasi-Newton method, conjugate gradients method (Fletcher and Reeves, 1964), the Nelder–Mead algorithm (Nelder and Mead, 1965) and SANN (Belisle, 1992). Hence if the default Nelder–Mead optimization method fails, the other methods can be used instead.
3. *No flexibility to choose alternative fit*: While the use of the method mentioned in this paper can provide some idea of a "definitive" fit, it does not allow the user to investigate other possible fits to the same data set. The use of numerical maximum log likelihood method is designed to give a general purpose fit rather than specific purpose fit. A more flexible method using bin-widths of the histogram that can accommodate a range of different fits on the same data can be found in Su (2005a).

6. Conclusion

This paper provides an alternative way of fitting distributions to data using two-step procedure, by first finding suitable initial values using method of moments or percentiles and then uses these initial values to maximize the numerical log likelihood. The results in this paper demonstrate that there are some differences in the result obtained between King and MacGillivray's (1999) goodness of fit method and numerical maximum log likelihood method. This highlights the importance of having a range of alternative fits to the same data, as there is currently neither a single method that will work the best in all cases, nor a consensus on a definitive method of fit. A typical trait in this type of research is that users often have to exercise judgment to choose the best alternatives for their purpose and perhaps future research on developing robust, general purpose diagnostic tools to assess the quality of fit may alleviate the users from this burden.

References

Au-Yeung, S., Dingle, N., Knottenbelt, W., 2004. Efficient approximation of response time densities and quantiles in stochastic models. Workshop on Software and Performance: Proceedings of the Fourth International Workshop on Software and Performance.

Belisle, C.P.J., 1992. Convergence theorems for a class of simulated annealing algorithms on R^d. J. Appl. Probab. 29, 885–895.

Bratley, P., Fox, B., 1988. Algorithm 659: implementing sobol's quasirandom sequence generator. ACM Trans. Math. Software 14 (1), 88–100.

Falta, M., Wolff, R., 2004. Recent developments of statistical approaches in aspects of accounting: a review. International Statistical Review 72 (3), 377–396.

Faure, H., Tezuka, S., 2000. Another random scrambling of digital (t, s)-sequences. MCQMC 2000.

Fletcher, R., Reeves, C.M., 1964. Function minimization by conjugate gradients. Comput. J. 7, 148–154.

Freimer, M., Mudholkar, G., Kollia, G., Lin, C., 1988. A study of the generalised Tukey lambda family. Commun. Statist. Theory Methods 17, 3547–3567.

Ganesalinggam, S., Kumar, K., 2001. Detection of financial distress via multivariate statistical analysis. Managerial Finance 27 (4), 45–55.

Hastings, J.C., Mosteller, F., Tukey, J., Winsor, C.W., 1947. Low moments for small samples: a comparative study of order statistics. Ann. Statist. 18, 413–426.

Hogben, D., 1963. Some Properties of Tukey's Test for Non-Additivity. Rutgers. The State University of New Jersey, New Brunswick.

Hong, H.S., Hickernell, F.J., 2002. Implementing scrambled digital sequences. Unpublished.

Karian, Z., Dudewicz, E., 2000. Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalised Bootstrap Methods. Chapman & Hall, New York.

Karian, Z., Dudewicz, E., 2003. Comparison of GLD fitting methods: superiority of percentile fits to moments in L^2 norm. J. Iranian Statist. Soc. 2 (2), 171–187.

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. Commun. Statist. Comput. Simulation 25 (3), 611–642.

King, R., MacGillivray, H., 1999. A starship estimation method for the generalised lambda distributions. Australia New Zealand J. Statist. 41 (3), 353–374.

Kumaran, M., Achary, K., 1996. On approximating lead time demand distributions using the generalised lambda type distribution. J. Oper. Res. Soc. 47 (4), 395–404.

Lakhany, A., Massuer, H., 2000. Estimating the parameters of the generalised lambda distribution. Algo Res. Quart. December, 47–58.

Nelder, J.A., Mead, R., 1965. A simplex algorithm for function minimization. Comput. J. 7, 308–313.

Okur, M., 1988. On fitting the generalised lambda distribution to air pollution data. Atmos. Environ. 22, 2569–2572.

Owen, A., 1995. Randomly permuted (t, m, s)-nets and (t, s)-sequences. in: Niederreiter, H., Shiue, P.J., (Eds.), Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, vol. 106. Springer, pp. 299–317.

Ozturk, A., Dale, R., 1985. Least squares estimation of the parameters of the generalised lambda distribution. Technometrics 27, 8–84.

Ramberg, J., Schmeiser, B., 1974. An approximate method for generating asymmetric random variables. Commun. ACM 17, 78–82.

Ramberg, J., Tadikamalla, P., Dudewicz, E., Mykytka, E., 1979. A probability distribution and its uses in fitting the data. Technometrics 21, 201–214.

Shapiro, S., Wilk, M., Chen, J.H., 1968. A comparative study of various tests of normality. J. Amer. Statist. Assoc. 63, 1343–1372.

Silverman, B.W., 1985. Density Estimation for Statistics and Data Analysis. Chapman & Hall, New York.

Su, S., 2005a. A discretized approach to flexibly fit generalized lambda distributions to data. J. Modern Appl. Statist. Methods, November, 408–424.

Su, S., 2005b. To match or not match? British Accounting Review March 37 (1), 1–21.

Wegman, E.J., 1972. Nonparametric probability density estimation. Technometrics 14, 533–546.

Willett, R., 1991. Transaction theory, stochastic processes and derived accounting measurement. Abacus September 1991(b), 117–134.