JIRSS (2003)
Vol. 2, No. 2, pp 171-187

Comparison of GLD Fitting Methods: Superiority of Percentile Fits to Moments in L^2 Norm

Zaven A. Karian[1], **Edward J. Dudewicz**[2]

[1]Department of Mathematics and Computer Science, Denison University, Granville, OH 43023. (karian@denison.edu)
[2]Department of Mathematics, Syracuse University, Syracuse, NY 13244. (dudewicz@mailbox.syr.edu)

Abstract. The flexibility of the family of Generalized Lambda Distributions (GLD) has encouraged researchers to fit GLD distributions to datasets in many circumstances. The methods that have been used to obtain GLD fits have also varied. This paper compares, for the first time, the relative qualities of three GLD fitting methods: the method of moments, a method based on percentiles, and a method that uses L-moments.

1 Introduction

The fitting of statistical distributions to data has broad applications and has been the subject many investigations. A central issue of such investigations is the choice of distribution or family of distributions to be fitted to the data. Because of its usefulness and flexibility in providing a model for data generated from scientific studies, the Gen-

Received: September 2003
Key words and phrases: Generalized lambda distribution, L-moments, method of moments, percentile method.

eralized Lambda Distribution, designated by GLD(λ_1, λ_2, λ_3, λ_4), and its predecessors have been studied by many researchers including: Tukey (1960), Ramberg and Schmiser (1972, 1974), Ramberg, Dudewicz, Tadikamalla and Mykytka (1979), Mykytka and Ramberg (1979), Dudewicz and Karian (1996, 1999), Öztürk and Dale (1985), Karian, Dudewicz and McDonald (1996), and Karian and Dudewicz (1999). The most recent and comprehensive is Karian and Dudewicz (2000).

The focus of much of the work cited above is on the determination of the GLD parameters in order to obtain the "best" fit to a given dataset. Several methods have been used in this effort: Karian, Dudewicz and McDonald (1996) use the method of moments (applied to the GLD and EGLD, an extension of the GLD system which covers all regions of (skewness, kurtosis) space), Mykytka (1976) uses a mixture of percentiles and moments to estimate the parameters of the GLD, Karian and Dudewicz (1999) use a percentile-based approach, Öztürk and Dale (1985) use least squares estimation, and Petersen (2001) uses a method based on L-moments.

As pointed out by King and MacGillivray (1999), "There ... appears to be a lack of assessment of any of the fitting methods. Thus, although the generalized λ distributions appear popular for use in simulation studies and appear to have considerable potential for fitting data, developing and assessing fitting methods for them is a challenge." The present paper addresses assessment quantitatively over a broad range of (skewness, kurtosis)-space, using methods not previously exploited for this purpose.

In Section 2 we define the GLD and EGLD and describe how to fit EGLD distributions to data using moments. Sections 3 and 4 develop methods for fitting GLD distributions using percentiles and L-moments, respectively. In subsequent sections we compare these GLD fitting methods by (1) applying these methods to datasets and comparing the p-values of chisquare goodness-of-fit tests associated with these fits, (2) examining the reliability of these p-values and (3) doing detailed comparisons of two of the fitting schemes (methods associated with moments and percentiles) at distinct (α_3, α_4) locations where α_3 and α_4 are the third and fourth central moments.

2 Fitting the GLD and EGLD via the method of moments

The GLD(λ_1, λ_2, λ_3, λ_4) family of distributions is defined through its quantile or inverse distribution function,

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

where $0 \leq y \leq 1$. This representation of $Q(y)$ is well-suited for simulation studies where, generally, random samples need to be generated. Karian and Dudewicz (2000) provide a complete analysis of the values of λ_1, λ_2, λ_3, λ_4 for which (1) defines a distribution (all λ_1, all λ_2 with the same sign that $\lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1-y)^{\lambda_4 - 1}$ has for all y ($0 < y < 1$), and some (λ_3, λ_4) regions are valid). The probability density function (*p.d.f.*) for the quantile function given in (1) is

$$f(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1-y)^{\lambda_4 - 1}}, \quad \text{at } x = Q(y). \tag{2}$$

The first two moments, the skewness and the kurtosis, respectively, of the GLD are given by (see Section 2.1 of Karian and Dudewicz (2000))

$$\begin{aligned}
\alpha_1 &= \mu = \lambda_1 + A/\lambda_2, && (3)\\
\alpha_2 &= \sigma^2 = (B - A^2)/\lambda_2^2, && (4)\\
\alpha_3 &= (C - 3AB + 2A^3)/(\lambda_2^3 \sigma^3), && (5)\\
\alpha_4 &= (D - 4AC + 6A^2 B - 3A^4)/(\lambda_2^4 \sigma^4), && (6)
\end{aligned}$$

where

$$\begin{aligned}
A &= 1/(1+\lambda_3) - 1/(1+\lambda_4),\\
B &= 1/(1+2\lambda_3) + 1/(1+2\lambda_4) - 2\beta(1+\lambda_3, 1+\lambda_4),\\
C &= 1/(1+3\lambda_3) - 1/(1+3\lambda_4) - 3\beta(1+2\lambda_3, 1+\lambda_4)\\
&\quad + 3\beta(1+\lambda_3, 1+2\lambda_4),\\
D &= 1/(1+4\lambda_3) + 1/(1+4\lambda_4) - 4\beta(1+3\lambda_3, 1+\lambda_4)\\
&\quad + 6\beta(1+2\lambda_3, 1+2\lambda_4) - 4\beta(1+\lambda_3, 1+3\lambda_4)
\end{aligned}$$

and $\beta(u, v)$ is the beta function given by

$$\beta(u, v) = \int_0^1 x^{u-1}(1-x)^{v-1}\, dx \text{ for } u, v > 0.$$

One way of fitting a specific GLD(λ_1, λ_2, λ_3, λ_4) to $X_1, X_2, \ldots, X_n$, a given dataset, is through the method of moments where we compute the first four sample moments, equate them to the first four GLD moments and solve the resulting equations for λ_1, λ_2, λ_3, λ_4. The first two sample moments are defined by

$$\hat{\alpha}_1 = \bar{X} = \sum_{i=1}^{n} X_i/n, \tag{7}$$

$$\hat{\alpha}_2 = \hat{\sigma}^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n, \tag{8}$$

and the sample skewness and kurtosis are defined, respectively, by

$$\hat{\alpha}_3 = \sum_{i=1}^{n} (X_i - \bar{X})^3/(n\hat{\sigma}^3), \tag{9}$$

$$\hat{\alpha}_4 = \sum_{i=1}^{n} (X_i - \bar{X})^4/(n\hat{\sigma}^4). \tag{10}$$

When these are set equal to their GLD counterparts, the complexity of the resulting equations forces us to seek numeric rather than closed-form solutions for λ_1, λ_2, λ_3, λ_4. Several authors have devised tabulated solutions for λ_1, λ_2, λ_3, λ_4 when (α_3^2, α_4) is within a "feasible" region; the most detailed tabulation of this sort is available in Appendix B of Karian and Dudewicz (2000).

There are some restrictions on the (α_3, α_4)-space that is covered by the GLD. In general, it is possible to have any (α_3, α_4) point that satisfies $1 + \alpha_3^2 < \alpha_4$, but the GLD fits are restricted to $1.8 + 1.7\alpha_3^2 < \alpha_4$, making the portion of (α_3, α_4)-space specified by $1 + \alpha_3^2 < \alpha_4 < 1.8 + 1.7\alpha_3^2$ unattainable through the method of moments. For this reason Karian, Dudewicz and McDonald (1996) devised an Extended GLD system, the EGLD, that uses a generalization of the beta distribution to cover the points of the region $1 + \alpha_3^2 < \alpha_4 < 1.8 + 1.7\alpha_3^2$. This Generalized Beta Distribution (GBD) is obtained by starting with a beta random variable, X, with p.d.f.

$$f(x) = \frac{\Gamma(\beta_3 + \beta_4 + 2)}{\Gamma(\beta_3 + 1)\Gamma(\beta_4 + 1)} x^{\beta_3}(1 - x)^{\beta_4}; \quad \beta_3, \beta_4 > -1, \quad 0 \le x \le 1 \tag{11}$$

defining the random variable Y by

$$Y = \beta_1 + \beta_2 X;$$

and showing that Y has p.d.f.

$$f(x) = C\beta_2^{-(\beta_3+\beta_4+1)}(x-\beta_1)^{\beta_3}(\beta_1+\beta_2-x)^{\beta_4}; \quad \beta_1 \le x \le \beta_1+\beta_2 \tag{12}$$

where

$$C = \frac{\Gamma(\beta_3+\beta_4+2)}{\Gamma(\beta_3+1)\Gamma(\beta_4+1)}. \tag{13}$$

Moreover, α_1, α_2, α_3, α_4 for the random variable Y are given by

$$\alpha_1 = \mu = \beta_1 + \beta_2(\beta_3+1)/B_2, \tag{14}$$

$$\alpha_2 = \sigma^2 = \frac{\beta_2^2(1+\beta_3)(1+\beta_4)}{B_2^2 B_3}, \tag{15}$$

$$\alpha_3 = \frac{2(\beta_4-\beta_3)\sqrt{B_3}}{B_4\sqrt{(\beta_3+1)(\beta_4+1)}}, \tag{16}$$

$$\alpha_4 = \frac{3B_3\left(\beta_3\beta_4 B_2 + 3\beta_3^2 + 5\beta_3 + 3\beta_4^2 + 5\beta_4 + 4\right)}{B_4 B_5(\beta_3+1)(\beta_4+1)}, \tag{17}$$

where

$$B_i = \beta_3 + \beta_4 + i \text{ for } i = 1, \ldots, 5.$$

Since it is reasonably difficult to solve the system of equations $\hat{\alpha}_i = \alpha_i$ $(i = 1, \ldots, 4)$, Dudewicz and Karian (1996) provide extensive tables for the numerical solutions of these equations, making it possible to obtain GBD fits either through these tables or by direct computation.

3 Fitting the GLD via a method based on percentiles

For a given dataset, $X_1, X_2, \ldots, X_n$, let $\tilde{\pi}_p$ denote the $(100p)$th percentile of the data. We compute $\tilde{\pi}_p$ by first writing $(n+1)p$ as $r+(a/b)$, where r is a positive integer and a/b is a fraction in the interval [0, 1). Then $\tilde{\pi}_p$ is obtained from the order statistics $Y_1, Y_2, \ldots, Y_n$ of the data by

$$\tilde{\pi}_p = Y_r + \frac{a}{b}\left(Y_{r+1} - Y_r\right)$$

(this definition of the $(100p)$th data percentile differs from definitions that are often used).

Karian and Dudewicz (1999) use the four sample statistics, $\tilde{\rho}_1$, $\tilde{\rho}_2$, $\tilde{\rho}_3$, $\tilde{\rho}_4$ defined by

$$\tilde{\rho}_1 = \tilde{\pi}_{0.5} \tag{18}$$

$$\tilde{\rho}_2 = \tilde{\pi}_{0.9} - \tilde{\pi}_{0.1} \tag{19}$$

$$\tilde{\rho}_3 = \frac{\tilde{\pi}_{0.5} - \tilde{\pi}_{0.1}}{\tilde{\pi}_{0.9} - \tilde{\pi}_{0.5}} \tag{20}$$

$$\tilde{\rho}_4 = \frac{\tilde{\pi}_{0.75} - \tilde{\pi}_{0.25}}{\tilde{\rho}_2} \tag{21}$$

to estimate the parameters of a GLD. The GLD counterparts of $\tilde{\rho}_1$, $\tilde{\rho}_2$, $\tilde{\rho}_3$, $\tilde{\rho}_4$ are

$$\rho_1 = Q(0.5) = \lambda_1 + \frac{0.5^{\lambda_3} - 0.5^{\lambda_4}}{\lambda_2} \tag{22}$$

$$\rho_2 = Q(0.9) - Q(0.1) = \frac{0.9^{\lambda_3} - 0.1^{\lambda_4} + 0.9^{\lambda_4} - 0.1^{\lambda_3}}{\lambda_2} \tag{23}$$

$$\rho_3 = \frac{Q(0.5) - Q(0.1)}{Q(0.9) - Q(0.5)} = \frac{0.9^{\lambda_4} - 0.1^{\lambda_3} + 0.5^{\lambda_3} - 0.5^{\lambda_4}}{0.9^{\lambda_3} - 0.1^{\lambda_4} + 0.5^{\lambda_4} - 0.5^{\lambda_3}} \tag{24}$$

$$\rho_4 = \frac{Q(0.75) - Q(0.25)}{\rho_2} = \frac{0.75^{\lambda_3} - 0.25^{\lambda_4} + 0.75^{\lambda_4} - 0.25^{\lambda_3}}{0.9^{\lambda_3} - 0.1^{\lambda_4} + 0.9^{\lambda_4} - 0.1^{\lambda_3}} \tag{25}$$

Since λ_1 can be any real value and $Q(\cdot)$ is an inverse distribution function, it follows from (22) through (25) that ρ_1, ρ_2, ρ_3, ρ_4 are subject to the constraints:

$$-\infty < \rho_1 < \infty, \qquad \rho_2 \geq 0, \qquad \rho_3 \geq 0, \qquad 0 \leq \rho_4 \leq 1.$$

To fit a GLD to a given dataset, we need to solve the system of equations $\tilde{\rho}_i = \rho_i$ $(i = 1, 2, 3, 4)$ for λ_1, λ_2, λ_3, λ_4.

Again, solutions cannot be attained in closed form and we use numerical methods. Appendix D of Karian and Dudewicz (2000) provides tables for estimating λ_1, λ_2, λ_3, λ_4 from $\tilde{\rho}_1$, $\tilde{\rho}_2$, $\tilde{\rho}_3$, $\tilde{\rho}_4$ in five distinct regions of (λ_3, λ_4)-space. It is quite likely that more than 1, possibly as many as 5, distinct GLD fits may be obtained.

4 Fitting the GLD via L-moments

Greenwood, Landwehr, Matalas and Wallis (1979) define the probability weighted moments of a random variable X by

$$M_{k,r,s} = E\left[X^k \left(F(X)\right)^r \left(1 - F(X)\right)^s\right]$$

where $F(X)$ is the distribution function of X. Let

$$\beta_j = M_{1,j,0} = E[X(F(X))^j]. \tag{26}$$

Then the L-moments of X are defined (see Hosking (1990)) as the linear combinations

$$\Lambda_1 = \beta_0, \quad \Lambda_i = \sum_{j=0}^{i-1} p_{i,j}\beta_j \quad \text{for} \quad i = 2, 3, \ldots \tag{27}$$

where

$$p_{i,j} = (-1)^{i-1-j}\binom{i-1}{j}\binom{i+j-1}{j} = \frac{(-1)^{i-1-j}(i+j-1)!}{(j!)^2(i-j-1)!}. \tag{28}$$

(Note that in the literature $L-$moments are usually denoted by λ_1, λ_2, λ_3, λ_4; we have chosen the Λ_1, Λ_2, Λ_3, Λ_4 notation here to avoid confusion with the λ_1, λ_2, λ_3, λ_4 of the GLD distribution.)

The first four L-moments, those of interest to us, are:

$$\begin{aligned}
\Lambda_1 &= \beta_0 && (29)\\
\Lambda_2 &= 2\beta_1 - \beta_0 && (30)\\
\Lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0 && (31)\\
\Lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0. && (32)
\end{aligned}$$

For the GLD the quantities Λ_1, Λ_2, Λ_3, Λ_4 calculate to

$$\Lambda_1 = \lambda_1 + \frac{1}{\lambda_2(\lambda_3+1)} - \frac{1}{\lambda_2(\lambda_4+1)} \tag{33}$$

$$\Lambda_2 = \frac{\lambda_3}{\lambda_2(\lambda_3+1)(\lambda_3+2)} + \frac{\lambda_4}{\lambda_2(\lambda_4+1)(\lambda_4+2)} \tag{34}$$

$$\begin{aligned}
\Lambda_3 &= \frac{\lambda_3(\lambda_3-1)}{\lambda_2(\lambda_3+1)(\lambda_3+2)(\lambda_3+3)}\\
&- \frac{\lambda_4(\lambda_4-1)}{\lambda_2(\lambda_4+1)(\lambda_4+2)(\lambda_4+3)}
\end{aligned} \tag{35}$$

$$\Lambda_4 = \frac{\lambda_3(\lambda_3-1)(\lambda_3-2)}{\lambda_2(\lambda_3+1)(\lambda_3+2)(\lambda_3+3)(\lambda_3+4)} + \frac{\lambda_4(\lambda_4-1)(\lambda_4-2)}{\lambda_2(\lambda_4+1)(\lambda_4+2)(\lambda_4+3)(\lambda_4+4)}. \quad (36)$$

To define the L-moments of a sample, we first take the order statistics of the sample: $x_1 \le x_2 \le \cdots \le x_n$ and define the sample L-moments, $\ell_1, \ell_2, \ldots, \ell_n$ by

$$\ell_i = \sum_{j=0}^{i-1} p_{i,j} b_j$$

where

$$b_j = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \frac{(i-1)(i-2)\cdots(i-j)}{(n-1)(n-2)\cdots(n-j)} x_i, \quad \text{for} \quad j = 0, 1, \ldots, n-1.$$

It is clear from (33) through (36) that the equations $\tau_3 = \Lambda_3/\Lambda_2 = l_3/l_2 = t_3$ and $\tau_4 = \Lambda_4/\Lambda_2 = l_4/l_2 = t_4$ will be free of λ_1 and λ_2, allowing us to solve them for λ_3 and λ_4. Once these values are obtained, equations (34) and (33) will yield λ_2 and λ_1, respectively. As was the case with percentile-based fits, multiple fits can be obtained through the use of L-moments.

5 Comparison of p-values

In this section we try to develop some insight into the relative merits of the three fitting schemes discussed in Sections 2, 3, and 4 by fitting GLD distributions to datasets through these methods and using the p-values of chi-square goodness-of-fit tests to determine the quality of each fit. These 13 datasets are the ones considered in Karian and Dudewicz (2000). (In Section 6 we look at two of these datasets in considerably greater detail.) Table 1 summarizes the results associated with fitting GLD distributions to datasets through the use of moments, percentiles and L-moments as well as GBD moment fits. The datasets are identified by the section numbers in Karian and Dudewicz (2000) in which they are discussed (e.g., one would find a discussion of the first dataset in Sections 2.5.2 and 3.5.3 of Karian and Dudewicz (2000)).

Table 1: p-Values Associated with Data Fits

Dataset	GLD Mom.	GLD Perc.	GLD L-Mom.	GBD Mom.
2.5.2, 3.5.3	.014	.022	**.076**	.020
2.5.4X	.725	.591 (92.5)	**.737**	*
2.5.4Y	.313	**.600**	.375	*
2.5.5X, 3.5.4X, 4.6.4X	.000	.180	*	**.290** (98.0)
2.5.5Y, 3.5.4Y, 4.6.4Y	*	**.062**	*	.009 (97.1)
2.6, 4.7	.686	.000 (79.2)	**.755**	*
3.5.1	*	*	*	**.353**
3.5.2	.063	.055	**.094**	.041
3.5.5X	.244	**.384** (82.1)	.336 (89.9)	.219
3.5.5Y	**.490**	.362 (91.1)	.476	.477 (94.1)
4.6.1	.000	**.590**	.335	*
4.6.2	*	**.104**	.0496	*
4.6.3	*	*	.000	.000 (76.1)

* Fit could not be obtained.

Regardless of the fitting method used, there is no assurance that the support of an EGLD distribution that has been fitted to X_1, X_2, $\ldots$, X_n will cover the span of the data (i.e., the interval $[\min_{1\le i\le n}(X_i), \max_{1\le i\le n}(X_i)]$). Failure to cover this span is designated in parentheses in Table 1. For example, for Data 3.5.5X, the GLD and GBD fits via moments covered the span of the data whereas the percentile and L-moment fits covered, respectively, 82.1% and 89.9% of the data span. Table 1 gives the p-values associated with the "best" fit when multiple fits are encountered. (Details of p-value calculation are given in Section 6.) By best, we mean the fit that covers the largest proportion of the data span and in case several fits produce the same proportional coverage, the one with the smallest χ^2 value (i.e., the largest p-value) is chosen. Within Table 1, the best fit for each dataset, the one with the largest p-value, is designated in boldface.

We note that there is at least one case where each of the four fitting methods is superior to all the others and, in general, the percentile and L-moment fits seem to give the best fits most frequently.

6 Comparison of moment and percentile fits

In this section we look, in some detail, at two datasets (3.5.5X and 2.5.4Y of Table 1) from Karian and Dudewicz (2000). Data 3.5.5X

represents measurements that are diameters of trees (in inches, at breast height) given by Schreuder and Hafley (1977). The actual data is:

3.30	5.90	3.80	7.00	5.80	4.40	4.60	8.20	3.30	10.20	11.30
4.80	9.10	4.40	5.60	8.50	8.20	4.70	8.00	8.20	7.70	10.00
4.80	7.10	5.50	12.80	8.90	9.80	7.10	6.10	10.00	8.10	2.20
7.30	10.80	5.60	6.40	4.30	9.50	7.70	7.40	3.60	5.50	6.50
10.30	4.90	5.80	14.80	6.40	4.90	10.00	3.40	6.30	8.90	8.10
10.30	3.00	2.40	4.70	4.10	4.00	4.20	10.70	4.00	2.50	10.20
3.50	9.30	8.60	9.10	8.10	6.90	5.50	5.80	5.50	10.40	4.40
4.70	4.70	4.50	4.90	3.00	10.30	5.30	8.80	7.80	6.50	7.20
7.20										

The α_1, α_2, α_3, α_4 for this data are:

$$6.7405,\ \ 6.6721,\ \ 0.45439,\ \ 2.7450$$

and the moment-based fit, designated by M.O.M in Table 2, is

$$\text{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)=\text{GLD}(L_0)=$$
$$\text{GLD}(4.774, 0.08911, 0.06257, 0.3056).$$

Using the 9 intervals

$$(-\infty, 3.75],\ (3.75, 4.55],\ (4.55, 5.05],\ (5.05, 6.05],\ (6.05, 7.05],$$
$$(7.05, 8.05],\ (8.05, 9.05],\ (9.05, 10.25],\ (10.25, \infty)$$

we see that the observed frequencies in these classes are:

$$10,\ 10,\ 10,\ 11,\ 8,\ 10,\ 11,\ 10,\ 9.$$

The expected frequencies of the chosen intervals above, based on the fitted GLD(4.774, 0.08911, 0.06257, 0.3056), are:
10.086, 9.1048, 6.5753, 13.596, 12.712, 10.905, 8.7720, 7.7285, 9.5195.
From this we are led to a chisquare goodness-of-fit statistic $\chi_0^2 = 5.452$ with a corresponding p-value of $p_0 = 0.2440$ (note that in this case we are estimating the four parameters λ_1, λ_2, λ_3, λ_4, therefore, the degrees of freedom of the relevant chisquare distribution is taken as $9 - 4 - 1 = 4$). In a similar fashion, we obtain χ_0^2 and p_0-values associated with percentile (designated by M.O.P in Table 2) and L-moment (designated by M.O.L. in Table 2) GLD fits. These are given in the first two rows, in columns 3, and 4 of Table 2.

Data 2.5.4Y is from the Indiana Twin Study and it consists of:[1]

[1]This particular data comes from the Ph.D. thesis of Dr. Cynthia Moore, under the supervision of Dr. Joseph C. Christian, Department of Medical and Molecular Genetics, Indiana University School of Medicine. The data collection was supported by the National Institutes of Health Individual Research Fellowship Grant: "Twin Studies in Human Development." PHS-5-F32-HD06869, 1987–1990.

2.81	3.78	2.93	4.13	3.19	5.38	3.56	3.24	3.16	3.83	3.81
4.60	3.66	4.28	5.00	4.75	6.31	4.31	4.75	4.50	3.66	3.88
4.69	3.40	5.00	5.13	4.15	3.83	4.31	5.38	4.12	4.63	4.56
4.63	4.38	4.44	5.13	4.78	4.22	5.38	6.16	4.94	3.81	5.00
5.75	6.69	4.94	4.75	5.38	5.63	5.81	6.10	6.25	4.69	5.69
6.81	5.69	5.75	5.13	6.75	5.38	4.44	5.48	6.31	6.22	6.18
4.69	4.88	4.97	5.38	5.06	5.63	5.19	5.94	5.56	4.88	5.69
5.88	5.88	5.50	4.81	5.41	5.75	6.31	5.63	5.31	5.19	6.13
5.85	4.44	5.50	5.81	6.10	8.14	5.19	6.05	6.38	6.16	6.53
6.19	6.19	6.38	7.10	5.81	6.13	6.56	6.22	6.19	6.60	7.41
6.06	6.63	5.50	5.72	7.31	8.00	4.58	7.31	7.22	7.25	6.44
7.75	6.31									

The α_1, α_2, α_3, α_4 for this data are:

5.3666, 1.2033, -0.012189, 2.7665

with moment-based fit

$$\text{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)=\text{GLD}(L_0)=$$
$$\text{GLD}(5.3904, 0.22933, 0.18838, 0.18073).$$

Using the 8 intervals

$$(-\infty, 4.00], (4.00, 4.65], (4.65, 5.05], (5.05, 5.45],$$
$$(5.45, 5.80], (5.89, 6.15], (6.15, 6.50], (6.50, \infty)$$

we see that the observed frequencies in these classes are:

15, 17, 16, 15, 15, 13, 16, 16.

The expected frequencies of the chosen intervals above, based on the fitted GLD(5.3904, 0.22933, 0.18838, 0.18073), are:

13.577, 18.617, 15.599, 17.179, 14.813, 13.280, 10.830, 19.106.

From this we are led to a chisquare goodness-of-fit statistic $\chi_0^2 = 3.558$ with a corresponding p-value of $p_0 = 0.3134$ (in this case we also are estimating the four parameters λ_1, λ_2, λ_3, λ_4, therefore, the degrees of freedom of the relevant chisquare distribution is $8-4-1 = 3$). In a similar fashion, we obtain χ_0^2 and p_0-values associated with percentile and L-moment GLD fits. These are given in the first two rows, in columns 6, and 7 of Table 2.

To analyze the reliability of the p_0-values associated with a fit, say the a GLD moment fit for Data 3.5.5X, we generate random samples, $X_1, X_2, \ldots X_k$ from GLD(L_0), find a moment-based fit for each random sample, and obtain χ_i^2, the chisquare statistic for this sample (using the same intervals that have already been established). We generate enough samples, k of them, to get 1000 cases where a fit is available. In Table 2 we see that for the data 3.5.5X moment-based fit, 1048 samples were required (48 is recorded as the number of failures to obtain a fit via moments). The row designated by $\chi_i^2 > \chi_0^2$

gives the number of instances (out of 1000) where $\chi_i^2 > \chi_0^2$. In the next two rows, we record the mean and variance, respectively, of $\chi_i^2, \chi_2^2, \ldots, \chi_{1000}^2$.

When fitting an actual dataset, one does not know the exact distribution from which it arises so to compare *methods*, one should sample from a *known* distribution. This is done in Section 7.

Table 2: p-Value Estimation

	3.5.5X			2.5.4Y		
	M.O.M	M.O.P	M.O.L	M.O.M	M.O.P	M.O.L
χ_0^2	5.452	4.162	4.556	3.558	1.870	3.113
p_0-value	.2440	.3845	.3360	.3134	.5998	.3746
Failures	48	334	7	7	22	0
$\chi_i^2 > \chi_0^2$	437	577	517	606	844	643
$\overline{\chi^2}$	6.032	5.653	5.237	5.788	4.432	4.542
$V(\chi^2)$	69.27	22.20	8.441	741.1	7.741	7.879

7 L^2-Norm comparison of moment and percentile fits

In this section we use L^2-norms to compare the qualities of moment and percentile fits at different (α_3, α_4) points. The L^2-norm of two functions $g(x)$ and $f(x)$ measures the discrepancy between $g(x)$ and $f(x)$ as

$$\int_{-\infty}^{\infty} |g(x) - h(x)|^2 \, dx.$$

Since GLD p.d.f.s cannot be expressed in closed form, the L^2-norm for two GLD p.d.f.s is evaluated numerically.

To obtain our comparison, we start by specifying 9 (α_3, α_4) points and at each point we

1. Obtain a GLD(λ_1, λ_2, λ_3, λ_4)=GLD(L_0), which has exactly (to numerical accuracy) $(0, 1, \alpha_3, \alpha_4)$; these are listed in Table 3,

2. Generate 1000 random samples, $X_1, \ldots, X_{1000}$, each of size 1000 from GLD(L_0) and for each such sample,

(a) Compute $(\alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \alpha_4^{(i)})$ and use it to obtain a GLD moment-based fit GLD(λ_1, λ_2, λ_3, λ_4)=GLD(LM_i)

(b) Compute $(\rho_1^{(i)}, \rho_2^{(i)}, \rho_3^{(i)}, \rho_4^{(i)})$ and use it to obtain a GLD percentile-based fit GLD(λ_1, λ_2, λ_3, λ_4)=(LP_i)

(c) Compute the L^2-norm, LM^2-N_i, of GLD(L_0) and GLD(LM_i)

(d) Compute the L^2-norm, LP^2-N_i, of GLD(L_0) and GLD(LP_i)

3. Determine the mean, $\overline{LM^2}$-N, of the LM^2-N_i and the mean, $\overline{LP^2}$-N, of the LP^2-N_i.

Table 3: $L_0 = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ for which GLD(L_0) has $(0, 1, \alpha_3, \alpha_4)$

(α_3, α_4)	$L_0 = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$
$A = (0.5, 7)$	$(0.4751, -4.4160, -5.1850, -1.4279)$
$B = (1, 9)$	$(-0.2148, -0.2355, -0.08436, -0.1249)$
$C = (1.5, 10.5)$	$(-0.2931, 0.2079, 12.1841, 66.0226)$
$D = (0.5, 10)$	$(-0.1207, 0.2231, 14.9626, 27.0019)$
$E = (1, 12)$	$(-0.1600, -0.3375, -0.1198, -0.1598)$
$F = (1.5, 14)$	$(-0.2595, -0.3099, -0.09815, -0.1591)$
$G = (0.5, 13)$	$(-0.06499, -0.41156, -0.1543, -0.1730)$
$H = (1, 15)$	$(-0.1327, -0.4036, -0.1413, -0.1790)$
$I = (1.5, 17)$	$(-0.2177, -0.3710, -0.1191, -0.1776)$

Even with the large sample size (1000) that is used, about 1% to 2% of samples are ones where moment-based fits cannot be obtained (this occurs in 146 out of the 9000 cases). The means, $\overline{LM^2}$-N given in Table 4 below, are the means of those situations where fits were possible. The reason for the failure to obtain fits on the 146 occasions is that the random sample produced an (α_3, α_4) point that was out of computation range.

Table 4: L^2-Norms of Moment and Percentile Fits

	Moment fits		Percentile fits	
$(\alpha_3,\ \alpha_4)$	$\overline{LM^2}$-N	No Fits	$\overline{LM^2}$-N	No Fits
$A = (0.5, 7)$	0.0307	4	0.0300	0
$B = (1, 9)$	0.0354	8	0.0308	0
$C = (1.5, 10.5)$	0.0425	21	0.0322	0
$D = (0.5, 10)$	0.0383	8	0.0305	0
$E = (1, 12)$	0.0425	13	0.0309	0
$F = (1.5, 14)$	0.0476	29	0.0316	0
$G = (0.5, 13)$	0.0447	11	0.0307	0
$H = (1, 15)$	0.0477	18	0.0301	0
$I = (1.5, 17)$	0.0515	34	0.0315	0

The quadratic least-squares fit of $\overline{LM^2}$-N from the moment-based fits to the 9 $(\alpha_3,\ \alpha_4)$ pairs is

$$\begin{aligned} f_M(\alpha_3, \alpha_4) &= .00898 - .00417\alpha_3 + .00381\alpha_4 + .00635\alpha_3^2 \\ &- .0000560\alpha_4^2 - .000548\alpha_3\alpha_4. \end{aligned}$$

The surface f_M, together with its percentile counterpart f_P, is shown in Figure 1. The predicted $\overline{LM^2}$-N values, $\widehat{\overline{LM^2}}$-$N$, from this least-squares fit at the 9 points are:

$$.0305, .0360, .0422, .0382, .0422, .0478, .0450, .0475, .0516$$

and the square of the correlation coefficient for $\overline{LM^2}$-N and $\widehat{\overline{LM^2}}$-$N$ is 0.998. The equation of the regression line is $y = 0.0000985 + 0.998x$ where x is $\overline{LM^2}$-N and y is $\widehat{\overline{LM^2}}$-$N$.

The quadratic least-squares fit of $\overline{LP^2}$-N from the percentile-based fits to the 9 $(\alpha_3,\ \alpha_4)$ pairs is

$$\begin{aligned} f_P(\alpha_3, \alpha_4) &= .0287 - .00235\alpha_3 + .000428\alpha_4 \\ &+ .00263\alpha_3^2 - .0000141\alpha_4^2 - .000120\alpha_3\alpha_4. \end{aligned}$$

The predicted $\overline{LP^2}$-N values, $\widehat{\overline{LP^2}}$-$N$, from this least-squares fit at the 9 points are:

$$.0301, .0306, .0322, .0305, .0307, .0318, .0306, .0305, .0313$$

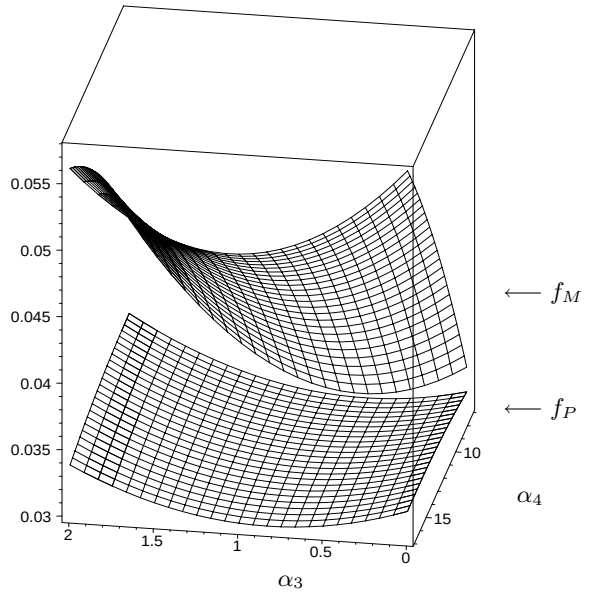

Figure 1. Approximations of L^2 norms assocated with moment (f_M) and percentile (f_P) fits.

and the square of the correlation coefficient for $\overline{LP^2}$-N and $\widehat{\overline{LP^2}}$-$N$ is 0.918. The equation of the regression line is $0.00254 + 0.918x$ where x is $\overline{LP^2}$-N and y is $\widehat{\overline{LP^2}}$-$N$.

From a comparison of the surfaces $f_M(\alpha_3\,\alpha_4)$ and $f_P(\alpha_3\,\alpha_4)$ in Figure 1 we conclude what Table 2 hinted at: *M.O.P. is superior to M.O.M. over a broad range of* (α_3, α_4)*-space* for fitting GLDs to samples of size 1000. The superiority is larger for larger α_4. In future studies we plan to extend these results to the "area between" norm given by $\int_{-\infty}^{\infty} |g(x) - f(x)|\,dx$, as well as to other methods, regions, and sample sizes.

References

Dudewicz, E. J. and Karian, Z. A. (1996), The extended generalized lambda distribution (EGLD) system for fitting distributions to data with moments. II: Tables, American Journal of Mathematical and Management Sciences, **16**(3 & 4), 271–332.

Dudewicz, E. J. and Karian, Z. A. (1999), Fitting the generalized lambda distribution (GLD) system by a method of percentiles. II: Tables, American Journal of Mathematical and Management Sciences, **19**(1 & 2), 1–73.

Greenwood, J. A., Landwehr, J. M., Matalas, M. C. and Wallis, J. R. (1979), Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. Water Resources Res., **15**, 1049–1054.

Hosking, J. R. M. (1990), *L*-Moments: Analysis and estimation of distributions using linear combinations of order statistics. J. Royal Stat. Soc., Ser. B, **52**, 105–124.

Karian, Z. A. and Dudewicz, E. J. (1999), Fitting the generalized lambda distribution to data: A method based on percentiles. Comm. Stat.: Simul. and Comput., **28** (3).

Karian, Z. A. and Dudewicz, E. J. (2000), Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods, CRC Press, Boca Raton, Florida.

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

King, R. A. R. and MacGillivray, H. L. (1999), A starship estimation method for the generalized lambda distributions. Australian and New Zealand Journal of Statistics, **41**(3), 353–374.

Mykytka E. F. (1976), Fitting a distribution to data using an alternative to moments. IEEE Proceedings of the 1976 Winter Simulation Conference, 361–374.

Öztürk, A. and Dale, R. F. (1985), Least squares estimation of the parameters of the generalized lambda distribution. Technometrics, **27**, 81–84.

Petersen, A. (2001), Personal communication based on his Master's Thesis. Frekvensanalyse av Hydrologisk og Meteorologisk Torke i Danmark (in Norwegian), Master's Thesis, Hovedoppgave ved Institutt For Geofysikk, University of Oslo, Norway.

Ramberg, J. S., Dudewicz, E. J., Tadikamalla, P. R., and Mykytka, E. F. (1979), A probability distribution and its uses in fitting data. Technometrics, **21**, 201–214.

Schreuder, H. T. and Hafley, W. L. (1977), A useful bivariate distribution for describing stand structure of tree heights and diameters. Biometrics, **33**, 471–478.