EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS
Vol. 7, No. 1, 2014, 22-36
ISSN 1307-5543 – www.ejpam.com

Empirical Likelihood Ratio Based Goodness-of-Fit Test for the Generalized Lambda Distribution

Wei Ning

Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, USA

Abstract. In this paper, we propose a goodness-of-fit test based on the empirical likelihood method for the generalized lambda distribution (GLD) family. Such a nonparametric test approximates the optimal Neyman-Pearson likelihood ratio test under the unknown alternative distribution scenario. The p-value of the test is approximated through the simulations due to the dependency of the test statistic on the data. The test is applied to the roller data set and the pollen data set to illustrate the testing procedure for the sufficiency of the GLD fittings.

2010 Mathematics Subject Classifications: 62G05, 62G10, 62G20

Key Words and Phrases: Generalized lambda distribution; Empirical likelihood; Nonparametric; Goodness-of-fit test.

1. Introduction

Modeling the skewed or the heavy tailed data is an important issue in statistical data fitting. There are extensive distribution families proposed by many researchers to achieve this goal. The generalized lambda distribution (GLD) family was originally introduced by Tukey [20], who proposed an one-parameter lambda distribution. Tukey's lambda distribution was generalized, for the purpose of generating random variables for Monte Carlo simulation studies, to the four parameters GLD proposed by Ramberg and Schmeiser [13, 14]. Ramberg *et al.* [15] developed a four-parameter system with the tables for fitting a wide variety of curve shapes. Since the early 1970s, the GLD has been applied to fitting phenomena in many fields of endeavor with continuous probability density functions (pdf). The GLD family with four parameters λ_1, λ_2, λ_3, λ_4, which is denoted as $GLD(\lambda_1,\lambda_2,\lambda_3,\lambda_4)$, has a probability density function

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

Email address: wning@bgsu.edu

where $Q(y)$ is the percentile function defined as

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

where $0 \leq y \leq 1$, λ_1 and λ_2 are location and scale parameters respectively. Karian and Dudewicz [4] gave the first four moments of $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ with $\lambda_3 > -1/4$, $\lambda_4 > -1/4$ as

$$\begin{aligned}
\alpha_1 &= \mu = E(X) = \lambda_1 + \frac{A}{\lambda_2}, \\
\alpha_2 &= \sigma^2 = E[(X-\mu)^2] = \frac{B - A^2}{\lambda_2^2}, \\
\alpha_3 &= E(X - E(X))^3/\sigma^3 = \frac{C - 3AB + 2A^3}{\lambda_3^3 \sigma^3}, \\
\alpha_4 &= E(X - E(X))^4/\sigma^4 = \frac{D - 4AC + 6A^2B - 3A^4}{\lambda_2^4 \sigma^4}
\end{aligned} \tag{2}$$

where

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

and $\beta(\cdot,\cdot)$ is a Beta function. The GLD family is known for its high flexibility on approaching many well-known distributions and ability to fit the data sets with different shapes, especially with those with heavy tails. There has been further extensive work done in this field. For example, Karian and Dudewicz [4] provided the tabulated tables for the method of moment and the percentile method which are used to estimate the parameters of the GLD. King and MacGillivray [6] and Lakhany and Massuer [7] considered a definite fit to the data set by maximizing the goodness of fit. Su [16] used a discretized method to fit GLD to the empirical data. Su [17] derived the estimation procedure by using the maximum likelihood method. Asquith [1] provided L-moments and TL-moments for the GLD. Fournier *et al.* [2] proposed a new estimation method by combining the method of moment and the percentile method. Ning *et al.* [8] considered the fitting problem involving the mixture of two GLDs and as a result made the comparisons to the other mixture distribution families. Ning and Gupta [9] proposed a GLD change point model to detect the change points for the DNA copy number. Su *et al.* [19] proposed a GLD based calibration model to achieve more flexible data fitting comparing the classic normal calibration model and the skew normal calibration model. For

the other recent work related to the GLD family and its applications, the readers are referred to Karian and Dudewicz [5].

As for data fitting, the extent to which how well the proposed distribution family can fit the data is an important issue. Goodness-of-fit test is the test that is always used to check the sufficiency of the data fitting by a given distribution. Neyman-Pearson lemma indicates that the likelihood ratio test (LRT) is the uniformly powerful (UMP) test for the hypotheses: $H_0 : f = f_0$ versus $H_1 : f = f_1$ with f_0 and f_1 both known. However, the alternative distribution is usually not known in practice. Recently, Vexler and Gurevich [22] and Vexler *et al.* [23] constructed goodness-of-fit test based on the empirical likelihood method to approximate the optimal Neyman-Pearson likelihood ratio test with an unknown alternative density function.

In this paper, we will consider the goodness-of-fit test for the generalized lambda distribution (GLD) family. We will adopt the idea as that of Vexler and Gurevich [22] similarly to construct a nonparametric goodness-of-fit test to test the null hypothesis of a GLD versus the alternative hypothesis of some other unknown distribution. This paper is organized as follows. In Section 2, a brief introduction of the empirical likelihood method will be given. The empirical likelihood ratio based goodness-of-fit test is proposed and its asymptotic properties will be derived. In Section 3, an empirical procedure based on the simulations is provided to approximate the p-value of the test statistic in Section 2 due to the dependency of the test statistic on the estimated parameters. The proposed goodness-of-fit test is applied to the roller data set and the pollen data set in Section 4 to illustrate the testing procedure and fitting results are given. Discussion is provided in Section 5.

2. Statistical Method

2.1. The Empirical Likelihood (EL) Method

Consider the independently and identically distributed p-dimensional observations, say $x_1, \cdots, x_n$, from an unknown population distribution F. The main idea of empirical likelihood methods, proposed and systematically developed by Owen [10] is to place a probability mass at each observation. Therefore, let $p_i = P(X = x_i)$ and the empirical likelihood function of F be defined as

$$L(F) = \prod_{i=1}^{n} p_i.$$

It is clear that $L(F)$ subject to the constraints

$$p_i \geq 0 \text{ and } \sum_i p_i = 1$$

is maximized at $p_i = 1/n$, i.e., the likelihood $L(F)$ attains its maximum n^{-n} under the full nonparametric model. When a population parameter θ identified by $Em(X, \theta) = 0$ is of interest where $m(x, \theta)$ is a real-valued function, the empirical log-likelihood maximum when θ has the true value θ_0 is obtained subject to the additional constraint

$$\sum p_i m(x_i, \theta_0) = 0.$$

The empirical log-likelihood ratio statistic to test $\theta = \theta_0$ is given by

$$R(\theta_0) = \max\{\sum_i \log np_i : p_i \geq 0, \sum p_i = 1, \sum p_i m(x_i, \theta_0) = 0\}.$$

Owen [10, 11] shows that similarly to the likelihood ratio test statistic in a parametric model setup, with mild regular conditions θ_0, $-2\log R(\theta_0) \to \chi^2_r$ in distribution under the null model $\theta = \theta_0$, where r is the dimension of $m(x, \theta)$. More details of the empirical likelihood and the related work refer to Owen [12].

2.2. The EL Goodness-of-Fit Test

We will test the following hypothesis:

$$H_0 : f = f_0 \sim GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$$
$$H_1 : f = f_1 \nsim GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4).$$

The likelihood ratio test statistic for this hypothesis is defined as

$$\text{LR} = \frac{\prod_{i=1}^n f_{H_1}(x_i)}{\prod_{i=1}^n f_{H_0}(x_i)} = \frac{\prod_{i=1}^n f_{H_1}(x_i)}{\prod_{i=1}^n f(x_i|\boldsymbol{\lambda})}$$

where $x_1, x_2, \cdots, x_n$ follows a GLD distribution with the parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ under the null hypothesis. Neyman-Pearson lemma guarantees that such a test is the UMP test with f_0 and f_1 both known. If they are both unknown, the maximum likelihood method will be applied to estimate the parameters $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$ of a GLD distribution under the null hypothesis. The parameters can be estimated by using the R package *GLDEX* developed by Su [18]. We then will apply maximum empirical likelihood method to estimate the numerator. We rewrite

$$L_f = \prod_{i=1}^n f_{H_1}(x_i) = \prod_{i=1}^n f_{H_1}(x_{(i)}) = \prod_{i=1}^n f_i,$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the order statistics of the observations $x_1, \cdots, x_n$. We will apply the empirical likelihood method introduced in Section 2.1 to derive the values of f_i to maximize L_f with the constraint $\int f(s)ds = 1$ corresponding to the alternative hypothesis. We first give the following lemma by Vexler and Gurevich [22] to express this constraint more explicitly.

Lemma 1. *Let $X_1, \cdots, X_n$ be independent and identically distributed random variables with a density function $f(x)$. Then*

$$\begin{aligned}\sum_{j=1}^n \int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx =& 2m\int_{X_{(1)}}^{X_{(n)}} f(x)dx - \sum_{k=1}^{m-1}(m-k)\int_{X_{(n-k)}}^{X_{(n-k+1)}} f(x)dx \\ &- \sum_{k=1}^{m-1}(m-k)\int_{X_{(k)}}^{X_{(k+1)}} f(x)dx \approx 2m\int_{X_{(1)}}^{X_{(n)}} f(x)dx - \frac{m(m-1)}{n},\end{aligned} \tag{3}$$

where $X_{(j)} = X_{(1)}$ *if* $j \leq 1$, *and* $X_{(j)} = X_{(n)}$, *if* $j \geq n$. $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ *are order statistics of* $X_1, \cdots, X_n$.

See Vexler and Gurevich [22] for the detailed proof.

Since $\int_{X_{(1)}}^{X_{(n)}} f(x)dx \leq \int_{-\infty}^{\infty} f(x)dx = 1$, from Lemma 1 we have

$$\Lambda_m \leq 1, \Lambda_m = \frac{1}{2m}\sum_{j=1}^{n}\int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx \tag{4}$$

Therefore, $\Lambda_m \to 1$ as $\frac{m}{n} \to \infty$. The integration on the right side of the equation (4) can be approximated as,

$$\sum_{j=1}^{n}\int_{X_{(j-m)}}^{X_{(j+m)}} f(x)dx \approx (X_{(j+m)} - X_{(j-m)})f(x_{(j)}) = (X_{(j+m)} - X_{(j-m)})f_j.$$

Thus,

$$\Lambda_m \approx \frac{1}{2m}\sum_{j=1}^{n}(X_{(j+m)} - X_{(j-m)})f_j \triangleq \widetilde{\Lambda}_m,$$

therefore, $\widetilde{\Lambda}_m \leq 1$. To maximize $\prod f_j$ with this constraint, we apply the Lagrange multiplier method and have

$$l(f_1, \cdots, f_n, \eta) = \sum_{j=1}^{n}\log f_j + \eta(\frac{1}{2m}\sum_{j=1}^{n}(X_{(j+m)} - X_{(j-m)})f_j - 1), \tag{5}$$

where η is a lagrange multiplier. By taking the derivative of the equation (5) respect to each f_j and η, we obtain

$$\frac{\partial l}{\partial f_i} = 0 \Rightarrow \frac{1}{f_j} + \frac{\eta}{2m}(X_{(j+m)} - X_{(j-m)}) = 0 \tag{6}$$

$$\frac{\partial l}{\partial \eta} = 0 \Rightarrow \frac{1}{2m}\sum_{j=1}^{n}(X_{(j+m)} - X_{(j-m)})f_j - 1 = 0. \tag{7}$$

from the equation (6) and (7), we have,

$$\sum f_j \cdot \frac{1}{f_j} + \eta\frac{1}{2m}\sum f_j(X_{(j+m)} - X_{(j-m)}) = 0 \Rightarrow \eta = -n. \tag{8}$$

Hence, we will obtain the estimate of f_j to maximize $\prod f_j$ as

$$f_j = \frac{2m}{n(X_{(j+m)} - X_{(j-m)})}, \tag{9}$$

where $X_{(j)} = X_{(1)}$, if $j \leq 1$, and $X_{(j)} = X_{(n)}$, if $j \geq n$. We then construct the likelihood ratio test statistic for the goodness-of-fit test for the GLD based on the maximum empirical likelihood method as

$$GLD_{mn} = \frac{\prod_{j=1}^{n} \frac{2m}{n(X_{(j+m)} - X_{(j-m)})}}{\max_{\boldsymbol{\lambda}} \prod_{j=1}^{n} f_{H_0}(X_j|\boldsymbol{\lambda})}, \tag{10}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ is the parameter vector of a GLD. We notice that the test statistic GLD_{mn} strongly depends on the integer m. To make the test more efficient, Vexler and Gurevich [22] and Vexler *et al.* [23] reconstructed the test statistic according to the properties of the empirical likelihood method. We follow their suggestions here to reconstruct the test statistic in (10) as

$$GLD_n = \frac{\min_{1 \leq m < n^{\delta}} \prod_{j=1}^{n} \frac{2m}{n(X_{(j+m)} - X_{(j-m)})}}{\max_{\boldsymbol{\lambda}} \prod_{j=1}^{n} f_{H_0}(X_j|\boldsymbol{\lambda})}, \tag{11}$$

and $0 < \delta < 1$. Here, we choose $\delta = 1/3$ for the convenience of computations. Then the equation (11) will be changed to

$$GLD_n = \frac{\min_{1 \leq m < n^{1/3}} \prod_{j=1}^{n} \frac{2m}{n(X_{(j+m)} - X_{(j-m)})}}{\prod_{j=1}^{n} f_{H_0}(X_j|\hat{\boldsymbol{\lambda}})}, \tag{12}$$

2.3. Asymptotic Results

In this section, we will derive some asymptotic properties of the test statistic proposed in (12). First we denote

$$h_i(x, \boldsymbol{\lambda}) = \frac{\partial \log f_{H_0}(x; \boldsymbol{\lambda})}{\partial \lambda_i}, i = 1, 2, 3, 4$$

and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. We assume the following conditions hold:

C1. $E(\log f(X_1))^2 < \infty$.

C2. Under the null hypothesis, $|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}| = \max_{1 \leq i \leq 4} |\hat{\lambda}_i - \lambda_i| \to 0$ in probability.

C3. Under alternative hypothesis, $\hat{\boldsymbol{\lambda}} \to \boldsymbol{\lambda}_0$ in probability where $\boldsymbol{\lambda}_0$ is a constant vector with finite components.

C4. There are open intervals $\Theta_0 \subseteq \mathbb{R}^4$ and $\Theta_1 \subseteq \mathbb{R}^4$ containing $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_0$ respectively.

There also exists a function $s(x)$ such that $|h(x, \boldsymbol{\xi})| \leq s(x)$ for all $x \in R$ and $\boldsymbol{\xi} \in \Theta_0 \cup \Theta_1$.

Theorem 1. *Assume that the conditions C1-C4 hold. Then under H_0,*

$$\frac{1}{n} \log(GLD_n) \to 0 \tag{13}$$

in probability as $n \to \infty$.

Theorem 2. *Assume that the conditions C1-C4 hold. Then under* H_1*,*

$$\frac{1}{n}\log(GLD_n) \to \mathrm{E}\log\left(\frac{f_{H_1}(X_1)}{f_{H_0}(X_1,\boldsymbol{\lambda}_0)}\right) \tag{14}$$

in probability as $n \to \infty$*. That is, the test is consistent.*

Please see both proofs in the Appendix.

3. Approximations to the p-value of GLD_n

From (12) in section 2.3, we observe that the values of the test statistic depend on the estimated parameters based on the data, and the asymptotic null distribution is not available. Therefore, we will provide an empirical procedure through the simulations to approximate the p-value asymptotically as follows.

1. Fit the original data $x_1, x_2, \ldots, x_n$ with a GLD and obtain the estimated values $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$.
2. Simulate a GLD distributed data $y_1, y_2, \cdots, y_n$ with the parameter $\hat{\boldsymbol{\lambda}}$.
3. Calculate the test statistic (12) for the original data and denote by GLD_n^1. Calculate the test statistic (12) for the simulated sample $y_1, \cdots, y_n$ and denote as GLD_n^{1B}.
4. Repeat the above simulation procedure M times and obtain M test statistics $GLD_n^{1B}, \cdots, GLD_n^{MB}$.
5. The p-value then will be approximated by

$$\hat{p} = \frac{1}{M}\sum_{i=1}^{M} I(GLD_n^{iB} \geq GLD_n^1),$$

where $I(\cdot)$ is an indicator function taking value 1 when $GLD_n^{iB} \geq GLD_n^1$, and taking value 0 when $GLD_n^{iB} < GLD_n^1$.

4. Application

In this section, we apply the proposed nonparametric goodness-of-fit test to two real data sets to show that GLD does offer sufficient fittings for both data sets.

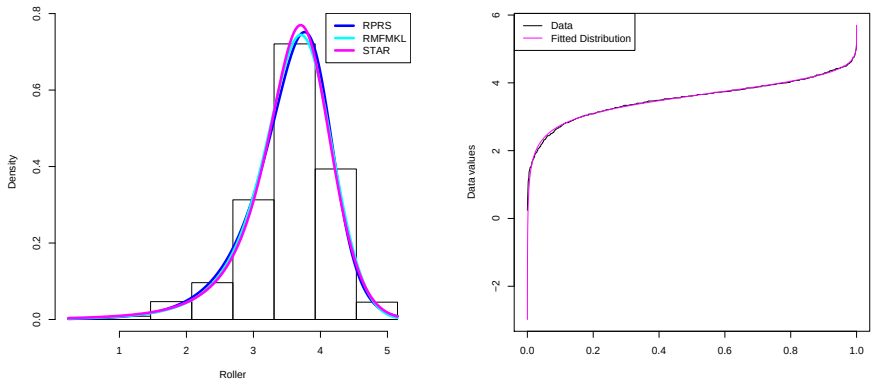
4.1. Roller Data

The data set consists of 1150 heights measured at 1 micron intervals along the drum of a roller for the purpose of the study of surface roughness of the rollers. The data is available at from Carnegie Mellon University Statistics Department*. We obtain the estimated parameters

*http://lib.stat.cmu.edu/jasadata/laslett

of an assumed GLD distribution of the data as $\hat{\lambda}_1 = 3.642$, $\hat{\lambda}_2 = 2.921$ and $\hat{\lambda}_3 = -0.145$, $\hat{\lambda}_4 = 0.166$ with the R package *GLDEX* by Su [18]. Following the procedure proposed in Section 3, we generate 1000 samples with $GLD(3.642, 2.921, -0.145, 0.166)$ with the sample size $n = 1150$. The test statistic is calculated from (12) with the approximated p-value as 0.068, which leads to fail to reject the null hypothesis at the significance level $\alpha = 0.05$, that is, the data can be fitted by a GLD model sufficiently.

We also apply the resample Kolmogorov-Smirnov test [Su 17] for the goodness-of-fit purpose. There are 960 times out of 1000 times that the p-value does not reject the null hypothesis, which indicates the sufficiency of the GLD fitting. Left graph in Figure 1 shows the fitted GLD density function with the histogram of the roller data. Three different estimated GLD density functions with different colors provided by GLDEX package [Su 17]. The dark blue one with the label RPRS corresponds to the GLD introduced in Section 2 due to Ramberg and Schmeiser [13], which is estimated by the maximum likelihood method. The light blue one with the label RMFMKL corresponds to a slight different version of GLD due to Freimer *et al.* [3], which is estimated by the maximum likelihood method. The pink one with the label STAR corresponds to the Freimer *et al.* [3] version of the GLD, which is estimated by the starship method [King and MacGillivray 6]). The right graph in Figure 1 is the quantile plot of the Ramberg and Schmeiser [13] version of the GLD fits. In Table 1, the estimated values of GLDs are all for Ramberg and Schmeiser [13] version of the GLD. From the graphs and tables, we can see the GLD fits the data pretty well. In Table 1, we also compare the first four moments of the real data with those of the fitted GLD distribution. From comparison we can observe that the moments of the fitted GLD are close to true moments of the data.



(a) Histogram of 1150 roller with GLD fits.

(b) Quantile plot for GLD fits of the data using maximum likelihood estimation.

Figure 1: Results of fitting GLD to the roller data.

Table 1: Moments comparison of roller data and the fitted GLD.

	DATA	GLD
Mean	3.53474	3.53544
Variance	0.42212	0.42135
Skewness	−0.98659	−0.98786
Kurtosis	4.86310	5.53253

4.2. Pollen Data

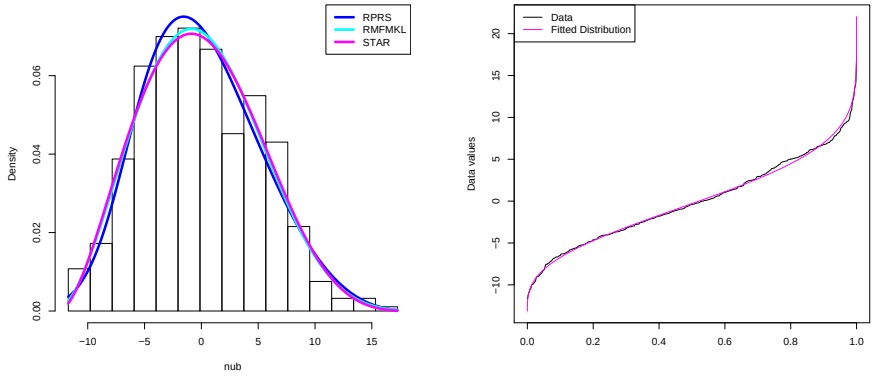
The second data is pollen data which also is available from Carnegie Mellon University[†]. We fit the variable "nub", which consists of 481 values from measuring geometric characteristics of certain type of pollen with a GLD model. We obtain the estimated parameters of an assumed GLD distribution of the data as $\hat{\lambda}_1 = -0.425$, $\hat{\lambda}_2 = 0.240$ and $\hat{\lambda}_3 = 0.328$, $\hat{\lambda}_4 = 0.186$ with the R package *GLDEX*. Following the procedure proposed in Section 3, we generate 1000 samples with $GLD(-0.425, 0.240, 0.328, 0.186)$ with the sample size $n = 481$. The test statistic is calculated from (12) with the approximated p-value as 0.156, which leads to fail to reject the null hypothesis at the significance level $\alpha = 0.05$, that is, the data can be fitted by a GLD model sufficiently. The results are listed in Table 2 provides the comparison of the moments of the data and the fitted GLD. We observe that the moments of the GLD model are close to true values of the moments of the data.

The resample Kolmogorov-Smirnov test [Su 18] shows that 959 times among 1000 times the p-value does not reject the null hypothesis, which indicates the sufficiency of the GLD fits. The histogram and quantile plots listed below also show the good fit of the GLD on the data. The left graph in Figure 2 shows the estimated density function of the GLDs by the maximum likelihood method and the starship method with the histogram of the pollen data. The colors have the same meaning as in the Figure 1. The right graph shows the quantile plot of the estimated Ramberg and Schmeiser [13] version of the GLD. We observe that the GLD fits the data very well.

Table 2: Moments comparison of pollen data and the fitted GLD

	DATA	GLD
Mean	−0.04826	−0.04813
Variance	26.94185	27.02739
Skewness	0.23295	0.25234
Kurtosis	2.59438	2.64018

[†]`http://lib.stat.cmu.edu/datasets/pollen.data`



(a) Histogram of 481 observations with GLD fits.

(b) Quantile plot for GLD fits of the data using maximum likelihood estimation.

Figure 2: Results of fitting GLD to the pollen data.

5. Discussion

In this paper, we investigate the problems of the goodness-of-fit for the generalized lambda distribution (GLD) family due to its high flexibility in the data fitting, especially for the heavy-tailed data. We propose an empirical likelihood based goodness-of-fit test for this distribution family. The motivation of the proposed test is based on the UMP likelihood ratio test guaranteed by Neyman-Pearson lemma with the null and alternative distributions both known. However, in practice, the alternative distribution is usually not known. Therefore, the proposed nonparametric goodness of fit test is constructed to approximate the optimal test for the scenario with an unknown alternative distribution. Asymptotic results have been derived for the proposed test. Since the explicit form of the asymptotic null distribution is not available and the test statistic is data dependent, we propose an empirical procedure through the simulations to approximate p-value of the test statistic for given data sets. The results show that the generalized lambda distribution offers sufficient fittings for both roller and pollen data sets, which match the conclusions obtained from the other goodness of fit tests such as the resample Kolmogorov-Smirnov test.

ACKNOWLEDGEMENTS The author would like to thank Professor Arjun Gupta for his valuable discussions and suggestions.

References

[1] W.H. Asquith. L-moments and TL-moments of the generalized lambda distribution. *Computational Statistics & Data Analysis*, 51:4484-4496, 2007.

[2] B. Fournier, N. Rupin, M. Bigerelle, D. Najjar, A. Iost and R. Wilcox. Estimating the parameters of a generalized lambda distribution. *Computational Statistics & Data Analysis*, 51:2813-2835, 2007.

[3] M. Freimer, G. Mudholkar, G. Kolloa. and C. Lin. A study of the generalized Tukey lambda family. *Communications in Statistics–Theory and Methods*, 17:3547-3567, 1988.

[4] Z.A. Karian and E.J. Dudewicz. *Fitting Statistical Distribution: The Generalized Lambda Distribution and Generalized Bootstrap Methods*, Boca Raton, FL:CRC Press, 2000.

[5] Z.A. Karian and E.J. Dudewicz. *Handbooks of Fitting Statistical Distributions with R*. New York: CRC Press, 2011.

[6] R. King and H. MacGillivray. A starship estimation methods for the generalized lambda distributions. *Australia & New Zealand Journal of Statistics*, 41:353-374, 1999.

[7] A. Lakhany and H. Massuer. Estimating the parameters of the generalized lambda distribution. *Algo Research Quarterly*, 47-58, 2000.

[8] W. Ning, Y.C. Gao and E.J. Dudewicz. Fitting Mixture Distributions Using Generalized Lambda Distributions and Comparisons with Normal Mixtures. *American Journal of Mathematical and Management Science*. Vol. 28, NOS. 1&2, 81-99, 2008.

[9] W. Ning and A.K. Gupta. Change point analysis for generalized lambda distributions. *Communication in Statistics-Simulations and Computation*. 38:1789-1802, 2009.

[10] A.B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237-249, 1988.

[11] A.B. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18:90-120, 1990.

[12] A.B. Owen. *Empirical Likelihood*. New York: CRC Press, 2001.

[13] J.S. Ramberg and B.W. Schmeiser. An approximation method for generating symmetric random variables. *Communications of the ACM*, 15: 987-990, 1972.

[14] J.S. Ramberg and B.W. Schmeiser. An approximation method for generating symmetric random variables. *Communications of the ACM*, 17:78-82, 1974.

[15] J.S. Ramberg, P.R. Tadikamalla, E.J. Dudewicz and E.F. Mykytka. A probability distribution and its uses in fitting data. *Technometrics*, 21:201-214, 1979.

[16] S. Su. A discretized approach to flexibility fit generalized lambda distributions to data. *Journal of Modern Applied Statistical Methods*, 4:402-424, 2005.

[17] S. Su. Numerical maximum log likelihood estimation for generalized lambda distribution. *Computational Statistics & Data Analysis*, 51:3983-3998, 2007.

[18] S. Su. Fitting single and mixture of generalized lambda distribution to data via dicretized and maximum likelihood methods: GLDEX in R. *Journal of Statistical Software*, 21:1-17, 2007.

[19] S. Su, A. Hasan. and W. Ning. (2013). The RS generalized lambda distribution based calibration model. *International Journal of Statistics and Probability*. 2:101-107, 2013.

[20] J.W. Tukey. The practice relationship between the common transformations of percentages of counts and of amounts. *Technical Report* 36, Statistical Techniques Research Group, Princeton University, 1960.

[21] O. Vasicek. A test for normality based on sample entropy. *Journal of Royal Statistical Society, B*, 38:54-59, 1976.

[22] A. Vexler and G. Gurevich. Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Computational Statistics and Data Analysis*, 54:531-545, 2010.

[23] A. Vexler, G. Shan, S.G. Kim, W.M. Tsai, L. Tian and A.D. Hutson. An empirical likelihood ratio based goodness-of-fit test for inverse Gaussian distributions. *Journal of Statistical Planning and Inference* 141:2128-2140, 2011.

Appendix

Proof. [Theorem 1] We consider the following statistic,

$$\begin{aligned} T_n =& \frac{1}{n}\log \min_{1\le m<n^{1/3}} \prod_{j=1}^{n} \frac{2m}{n(X_{(j+m)}-X_{(j-m)})} \\ =& -\max_{1\le m<n^{1/3}} t_{mn} \end{aligned}$$

where $t_{mn} = \frac{1}{n}\sum_{i=1}^{n}\log(\frac{n}{2m}(X_{(j+m)}-X_{(j-m)})$. It is a part of $n^{-1}\log(GLD_n)$, where GLD_n is the test statistic defined in (12). Similar to the work by Vasicek [21], we rewrite

$$t_{mn} = -\frac{1}{n}\sum_{i=1}^{n}\log f(x_i) + V_{mn} + U_{mn}, \tag{A1}$$

where

$$\begin{aligned} V_{mn} =& -\frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{F(x_{(i+m)})-F(x_{(i-m)})}{f(x_{(i)})(x_{(i+m)}-x_{(i-m)})}\right], \\ U_{mn} =& \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{n}{2m}(F(X_{(i+m)})-F(X_{(i-m)}))\right]. \end{aligned}$$

where F is the distribution function of $X's$. Denote F_n the empirical distribution function of $X's$ and combine the first two terms in (A-1), we obtain

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^{n}\log f(x_i) + V_{mn} =& -\frac{1}{n}\sum_{i=1}^{n}\log f(x_i) - \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{F(x_{(i+m)})-F(x_{(i-m)})}{f(x_{(i)})(x_{(i+m)}-x_{(i-m)})}\right] \\ =& -\frac{1}{n}\sum_{i=1}^{n}\log f(x_{(i)}) - \frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{F(x_{(i+m)})-F(x_{(i-m)})}{f(x_{(i)})(x_{(i+m)}-x_{(i-m)})}\right] \\ =& -\frac{1}{n}\sum_{i=1}^{n}\log\left[\frac{F(x_{(i+m)})-F(x_{(i-m)})}{(x_{(i+m)}-x_{(i-m)})}\right] \\ =& -\frac{1}{2m}\sum_{j=1}^{2m}\sum_{i=1}^{n}\log\left[\frac{F(x_{(i+m)})-F(x_{(i-m)})}{(x_{(i+m)}-x_{(i-m)})}\right](F_n(x_{(i+m)})-F_n(x_{(i-m)})), \end{aligned}$$

where $i \equiv j \pmod{2m}$. Let

$$S_j = -\sum_{i=1}^{n}\log\left[\frac{F(x_{(i+m)})-F(x_{(i-m)})}{(x_{(i+m)}-x_{(i-m)})}\right](F_n(x_{(i+m)})-F_n(x_{(i-m)})), i \equiv j \pmod{2m},$$

then

$$t_{mn} = \frac{1}{2m}\sum_{j=1}^{2m} S_j + U_{mn}.$$

With the argument in Theorem 1. by Vasicek [21],

$$\frac{1}{2m}\sum_{j=1}^{2m} S_j \to H(f), \quad a.s,$$

as $m/n \to 0$ uniformly for all $1 \le m \le n^{1/3}$, where $H(f) = E(-\log f(x_i)) = E(-\log f(x_1)$. Since the statistics U_{mn} is a non-positive variable by the definition and is independent of F, $U_{mn} \to 0$ in probability as $m \to \infty$ and $n \to \infty$ by Lemma 1 in Vasicek [21]. Therefore,

$$\begin{aligned} T_n &\le -\, t_{n^{1/3},n} \xrightarrow{p} E_f(\log(X_1)), \\ T_n &\ge -\max_{1\le m<n^{1/3}} (2m)^{-1} \sum_{j=1}^{2m} S_j \xrightarrow{p} E_f(\log(X_1)) \end{aligned}$$

as $n \to \infty$. It implies that

$$T_n \to E_f(\log(f(X_1))), \text{ as } n \to \infty. \tag{A2}$$

Now, we rewrite the test statistic GLD_n as

$$\begin{aligned} \frac{1}{n}\log(GLD_n) =& T_n - \frac{1}{n}\sum_{i=1}^{n}\log(f_{H_0}(X_i|\boldsymbol{\lambda})) \\ &+ \frac{1}{n}\left(\sum_{i=1}^{n}\log(f_{H_0}(X_i|\boldsymbol{\lambda})) - \sum_{i=1}^{n}\log(f_{H_0}(X_i|\hat{\boldsymbol{\lambda}}_n))\right), \end{aligned} \tag{A3}$$

where $\hat{\boldsymbol{\lambda}}_n = (\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$. Under the null hypothesis, $T_n \to E_{f_{H_0}}(\log(f_{H_0}(X_1)))$ in probability since (A2). With the condition (**C1**), we have

$$\frac{1}{n}\sum_{i=1}^{n}\log(f_{H_0}(X_i|\boldsymbol{\lambda})) \xrightarrow{p} E_{f_{H_0}}(\log(f_{H_0}(X_1|\boldsymbol{\lambda}))). \tag{A4}$$

With the condition (**C2**) holds and applying one-term Taylor Series expansion to the third part in the equation (A3), we obtain

$$\frac{1}{n}\left[\sum_{i=1}^{n}\log(f_{H_0}(X_i|\boldsymbol{\lambda})) - \sum_{i=1}^{n}\log(f_{H_0}(X_i|\hat{\boldsymbol{\lambda}}_n))\right] \cong \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{4} h_i(X_i;\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}})(\theta_j - \hat{\lambda}_{nj}),$$

where $h_i(\cdot)$ is defined in Section 2.3. Since (**C4**), we obtain

$$\begin{aligned} \frac{1}{n}\{\sum_{i=1}^{n}\log(f_{H_0}(X_i|\boldsymbol{\lambda})) - \sum_{i=1}^{n}\log(f_{H_0}(X_i|\hat{\boldsymbol{\lambda}}_n))\} &\cong \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{4} h_i(X_i;\hat{\boldsymbol{\lambda}}_{\boldsymbol{n}})(\lambda_j - \hat{\lambda}_{nj}) \\ &\ge \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{4} |h_i(X_i;\xi_i)|(\lambda_j - \hat{\lambda}_{nj}) \xrightarrow{p} 0 \end{aligned} \tag{A5}$$

where $|\xi_i - \boldsymbol{\lambda}| \leq |\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_n|$. Thus under the null hypothesis, the equations (A3), (A4) and (A5) provide

$$\frac{1}{n}\log(GLD_n) \xrightarrow{p} 0, \text{ as } n \to \infty. \tag{A6}$$

This completes the proof of Theorem 1. □

Proof. [Theorem 2] Under H_1, we have

$$\begin{aligned}\frac{1}{n}\log(GLD_n) =& T_n - \frac{1}{n}\sum_{i=1}^{n}\log(f_{H_1}(X_i)) + \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{H_1}(X_i)}{f_{H_0}(X_i|\boldsymbol{\lambda}_0)}\right) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{H_0}(X_i|\boldsymbol{\lambda}_0)}{f_{H_0}(X_i|\hat{\boldsymbol{\lambda}}_n)}\right)\end{aligned} \tag{A7}$$

Since (A2), similarly under H_1

$$T_n \to E_{f_{H_1}}(\log(f_{H_1}(X_1))), \text{ as } n \to \infty, \tag{A8}$$

and the condition (**C1**) leads to

$$\frac{1}{n}\sum_{i=1}^{n}\log(f_{H_1}(X_i|\boldsymbol{\lambda})) \xrightarrow{p} E_{f_{H_1}}(\log(f_{H_1}(X_1|\boldsymbol{\lambda}))). \tag{A9}$$

With the condition (**C3**),

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{H_0}(X_i|\boldsymbol{\lambda}_0)}{f_{H_0}(X_i|\hat{\boldsymbol{\lambda}}_n)}\right) \xrightarrow{P} 0 \tag{A10}$$

as $n \to 0$. With the equations (A8),(A9) and (A10), we obtain,

$$\frac{1}{n}\log(GLD_n) \xrightarrow{P} \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{H_1}(X_i)}{f_{H_0}(X_i|\boldsymbol{\lambda}_0)}\right) > 0 \text{ as } n \to \infty. \tag{A11}$$

□