Change Point Analysis for Generalized Lambda Distribution

Article *in* Communication in Statistics- Simulation and Computation · December 2009

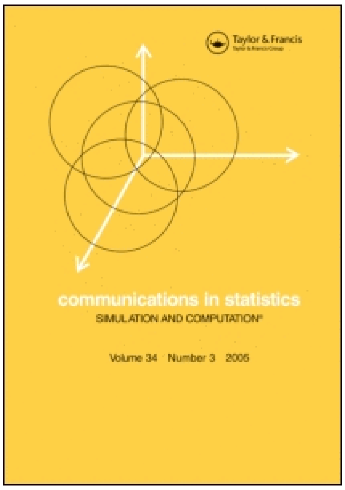
DOI: 10.1080/03610910903125314 · Source: DBLP

Wei Ning
Bowling Green State University
68 PUBLICATIONS 234 CITATIONS

SEE PROFILE

Arjun Gupta
Bowling Green State University
531 PUBLICATIONS 7,821 CITATIONS

SEE PROFILE



Communications in Statistics - Simulation and Computation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597237

Change Point Analysis for Generalized Lambda Distribution

Wei Ning[a]; A. K. Gupta[a]
[a] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio, USA

Online publication date: 09 December 2009

To cite this Article Ning, Wei and Gupta, A. K.(2009) 'Change Point Analysis for Generalized Lambda Distribution', Communications in Statistics - Simulation and Computation, 38: 9, 1789 — 1802

To link to this Article: DOI: 10.1080/03610910903125314

URL: http://dx.doi.org/10.1080/03610910903125314

Communications in Statistics—Simulation and Computation®, 38: 1789–1802, 2009
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910903125314

Change Point Analysis for Generalized Lambda Distribution

WEI NING AND A. K. GUPTA

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

In this article, we study the detection of multiple change points of parameters of generalized lambda distributions (GLD). The advantage of studying GLD is that the GLD family is broad and flexible. Compared to the other distributions, there are fewer restrictions on the distribution while fitting data. We combine the binary segmentation procedure together with the Schwarz information criterion (SIC) to search for all possible change points in the data. The method is applied on fibroblast cancer cell line data which is publicly available, and the change points are successfully located.

Keywords Change points; Estimation; Generalized lambda distributions; Information criterion.

Mathematics Subject Classification 60F10; 62N01; 62N02; 62F10; 62F12; 62F15.

1. Introduction

1.1. *The Change Point Problem*

Change point problems can be encountered in many applied fields such as finance, biology, geology etc., and even in our daily lives. In statistics, a change point can be viewed as a place or time point such that the observations before that point follow one distribution, and follow another distribution after that point. Multiple change points can be defined similarly. So the change point analysis usually gets involved in two problems: one is the existence of any change point among the data and the other is the detection of the change point if there is a change point. Many researchers have contributed to this field since the earliest change point study in the 1950s. For instance, Chernoff and Zacks (1964), Gardner (1969), Hawkins (1992), Sen and Srivastava (1975), and Worsley (1979) studied the testing and estimation of a change in the mean of a normal model. Hsu (1977), Inclán (1993), Wichern et al. (1976) studied change point problem for the variance. Hsu (1979) studied the

Received December 9, 2008; Accepted June 16, 2009

Address correspondence to Wei Ning, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA; E-mail: wning@bgsu.edu

shift of parameter in gamma models. Worsley (1986) provided confidence regions and tests for a change-point in a sequence of exponential models. The change point problem for the regression model has been studied by Krishnaiah and Miao (1988). Kim and Siegmund (1989) proposed a likelihood ratio test to detect a single change point in a simple linear regression model. Chen and Gupta (1997) studied the change points for the variance while the mean is constant for univariate normal model using information approach. Chen and Gupta (2000) extended their results to the multivariate normal models. They also discussed the testing and detection of change points for some continuous distributions besides the normal distribution, such as the exponential distributions, and also for some discrete distributions such as the gamma distribution and the binomial distributions by using likelihood ratio test (LRT), Bayesian approach and information approach (see Chen and Gupta, 2000). Hartigan (1990) introduced the product partition model to combine data from different sources and applied it to fatalities in manned rocket launches. Barry and Hartigan (1993) conducted a Baysian analysis for change point problems using the product partition model, and demonstrated that the proposed model was superior to the other alternatives in detecting sharp short-lived changes in the parameters. Loschi et al. (2005) studied the multiple change point problem for the regular exponential family using product partition model.

1.2. *The Generalized Lambda Distribution*

Pearson (1895) gave a four-parameter system of probability density functions, and fitted the parameters by the method of moments (MME). Tukey (1960) proposed a one-parameter lambda distribution. Tukey's lambda was generalized, for the purpose of generating random variables for Monte Carlo simulation studies, to the four-parameter generalized lambda distribution (GLD), by Ramberg and Schmeiser (1972, 1974). Ramberg et al. (1979) developed a four-parameter model together with the necessary tables for fitting a wide variety of curves. Since the early 1970s, the GLD has been applied in many fields of endeavor with continuous probability density functions. In this section, we briefly introduce the generalized lambda distribution and its properties. For more details about the GLD, see Karian and Dudewicz (2000).

The generalized lambda distribution family with four parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, denoted by GLD($\lambda_1, \lambda_2, \lambda_3, \lambda_4$), has the density function

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

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

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

where $0 \leq y \leq 1$. Here, λ_1 and λ_2 are the location and scale parameters, respectively, and λ_3 and λ_4 determine the skewness and kurtosis of the GLD($\lambda_1, \lambda_2, \lambda_3, \lambda_4$). Not all choices of $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ lead to a valid distribution, as described in the following theorem.

Theorem 1.1. *The* $\mathrm{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ *specifies a valid distribution if and only if*

$$g(y, \lambda_3, \lambda_4) \equiv \lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1 - y)^{\lambda_4 - 1} \tag{2}$$

has the same sign for all y in $[0, 1]$, *as long as* λ_2 *takes that sign also. In particular, the* $\mathrm{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ *specifies a valid distribution if* $\lambda_2, \lambda_3, \lambda_4$ *all have the same sign.*

The first four moments of $\mathrm{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are given as follows.

Theorem 1.2. *If X is* $\mathrm{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ *with* $\lambda_3 > -1/4$ *and* $\lambda_4 > -1/4$, *then its first four moments,* $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, *are given by*

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

where

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

As a consequence of Theorem 1.2, the following theorem indicates that a random variable with a distribution other than the GLD can be approximated by a GLD with some $\lambda_1, \lambda_2, \lambda_3, \lambda_4$.

Theorem 1.3. *For a given random variable Y which has a distribution other than the GLD, and its first four moments are* $\alpha_1 = \mu$, $\alpha_2 = \sigma^2$, α_3, *and* α_4, *it can be approximated by a random variable X that is* $\mathrm{GLD}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ *if we can choose* λ_3 *and* λ_4 *so that a* $\mathrm{GLD}(0, 1, \lambda_3, \lambda_4)$ *has the third and fourth moment* α_3 *and* α_4, *then* λ_1 *and* λ_2 *are the solutions of the equations*

$$\mu = \lambda_1 + \frac{A}{\lambda_2}, \quad \sigma^2 = \frac{B - A^2}{\lambda_2^2},$$

where A, B are defined in Theorem 1.2.

Karian and Dudewicz (2000, P_{48}), gave more details of the (α_3^2, α_4)-space associated with the GLD$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. The generalized lambda distribution family is broad since it can approximate many discrete distributions as well as continuous distributions. For example, GLD(0, 0.1975, 0.1349, 0.1349) can approximate $N(0, 1)$ with the error 0.001085; GLD(0.5, 2.0, 1.0, 1.0) can approximate $U([0, 1])$ perfectly, etc. (see Karian and Dudewicz, 2000, Ch. 2, for more examples).

The purpose of this article is to use information approach for the GLDs to detect possible change points for a given data set. The article is organized as follows. Section 2 will give the hypotheses and information approach procedure corresponding to the change point problems for GLDs. Section 3 derives the estimations of the parameters under the null and alternative hypotheses, respectively. In Sec. 4, to avoid the effects of the random noise from the data, we introduce the critical values c_α with different significance levels α to make our conclusions about the change points more statistically convincing. Simulations are conducted in Sec. 5 to verify that the test procedure is powerful to detect the change points at different locations in a data. We also compare the power of the testing procedure using the normal distribution. The simulation results indicate that the GLD change point model is more powerful than the normal change point model on detecting the change points for a non symmetric data. The method is applied to the cell line, GM01750, one of 15 fibroblast cancer cell lines analyzed by Snijders et al. (2001) in Sec. 6. Discussion is provided in Sec. 7.

2. Information Approach

Let $X_1, X_2, \dots, X_n$ be a sequences of independent random variables with generalized lambda distributions with parameters

$$\left(\lambda_1^{(1)}, \lambda_2^{(1)}, \lambda_3^{(1)}, \lambda_4^{(1)}\right), \quad \left(\lambda_1^{(2)}, \lambda_2^{(2)}, \lambda_3^{(2)}, \lambda_4^{(2)}\right), \dots, \left(\lambda_1^{(n)}, \lambda_2^{(n)}, \lambda_3^{(n)}, \lambda_4^{(n)}\right).$$

Suppose that each x_i has a generalized lambda distribution with $\lambda_1^{(i)}, \lambda_2^{(i)}, \lambda_3^{(i)}$, and $\lambda_4^{(i)}$. In general, we would like to test the hypotheses are

$$\begin{aligned} &H_0 : \lambda_j^{(1)} = \lambda_j^{(2)} = \cdots = \lambda_j^{(n)} = \lambda_j \\ &H_1 : \lambda_j^{(1)} = \cdots = \lambda_j^{(k_1)} \neq \lambda_j^{(k_1+1)} = \cdots = \lambda_j^{(k_2)} \neq \cdots \neq \lambda_j^{(k_q+1)} = \cdots \lambda_j^{(n)}, \end{aligned} \tag{4}$$

where $j = 1, 2, 3, 4$ and q is the unknown number of change points and $1 < k_1 < k_2 < \cdots < k_q < n$ are unknown positions of the change points. This is the typical multiple change point problem discussed by Chen and Gupta (2000). To detect the number of change points in a multidimensional random process, Vostrikova (1981) proposed a binary segmentation method. This method has the advantages of detecting the number of change points and their positions simultaneously and of saving a lot of computation time. In this procedure, we first detect a single change at the first stage. If there is no change, we accept H_0. If there is a change, then such a change point divides the original sequence of random variables into two subsequences. For each subsequence, repeat the detection procedure in the first stage, and continue such a process until no more changes are found in any of the subsequences.

Based on the binary segmentation method, we just need to test the single change point hypothesis and repeat the process for each subsequence until the null hypothesis is accepted. Therefore, we turn to test H_0 against the following alternative hypothesis

$$H_1 : \lambda_j^{(1)} = \cdots = \lambda_j^{(k)} \neq \lambda_j^{(k+1)} = \cdots = \lambda_j^{(n)}, \quad j = 1, 2, 3, 4 \tag{5}$$

where $1 < k < n$ is the unknown position of the change point. Schwarz Information Criterion is expressed as

$$\mathrm{SIC}_p = -2 \cdot \log L(\widehat{\Theta}_p) + p \cdot \log n, \quad p = 1, 2, \ldots, K,$$

where K is the number of parameters of the model. Schwarz Information Criterion under the null hypothesis is defined as

$$\mathrm{SIC}(n) = -2\left[n \cdot \log \hat{\lambda}_2 - \sum_{i=1}^{n} \log\left(\hat{\lambda}_3 y_i^{\hat{\lambda}_3 - 1} + \hat{\lambda}_4 (1 - y_i)^{\hat{\lambda}_4 - 1}\right)\right] + 4 \cdot \log n.$$

Schwarz Information Criterion under the alternative hypothesis is defined as

$$\mathrm{SIC}(k) = -2\Bigg[k \cdot \log \hat{\lambda}_2^{(1)} + (n-k) \cdot \log \hat{\lambda}_2^{(n)} - \sum_{i=1}^{k} \log\left(\hat{\lambda}_3^{(1)} y_i^{\hat{\lambda}_3^{(1)} - 1} + \hat{\lambda}_4^{(1)} (1 - y_i)^{\hat{\lambda}_4^{(1)} - 1}\right)$$
$$- \sum_{i=k+1}^{n} \log\left(\hat{\lambda}_3^{(n)} y_i^{\hat{\lambda}_3^{(n)} - 1} + \hat{\lambda}_4^{(n)} (1 - y_i)^{\hat{\lambda}_4^{(n)} - 1}\right)\Bigg] + 8 \cdot \log n.$$

Note that to be able to obtain the maximum likelihood estimators, we can only detect changes for $2 \leq k \leq n-2$. Hence, according to the principle of information criterion, we do not reject H_0 if

$$\mathrm{SIC}(n) < \min_{2 \leq k \leq n-2} \mathrm{SIC}(k),$$

and accept H_1 if

$$\mathrm{SIC}(n) > \mathrm{SIC}(k)$$

for some k. We estimate the position of the change point by $\hat{k}$ such that

$$\mathrm{SIC}(\hat{k}) = \min_{2 \leq k \leq n-2} \mathrm{SIC}(k).$$

To calculate $\mathrm{SIC}(n)$ and $\mathrm{SIC}(k)$, we need to find the MLEs of the parameters under H_0 and H_1.

3. Estimation of Parameters

Under H_0, the likelihood function is

$$L_0 = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{\lambda_2}{\lambda_3 y_i^{(\lambda_3 - 1)} + \lambda_4 (1 - y_i)^{(\lambda_4 - 1)}}. \tag{6}$$

The log likelihood function is

$$\log L_0 = n \cdot \log \lambda_2 - \sum_{i=1}^{n} \log\left(\lambda_3 y_i^{(\lambda_3-1)} + \lambda_4 (1-y_i)^{(\lambda_4-1)}\right), \tag{7}$$

where

$$x_i = Q(y_i) = \lambda_1 + \frac{y_i^{\lambda_3} - (1-y_i)^{\lambda_4}}{\lambda_2} \tag{8}$$

and $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. Based on log likelihood function, the MLEs of the parameters are obtained from the following equations:

$$\frac{\partial \log L_0}{\partial \lambda_1} = -\sum_{i=1}^{n} \lambda_2 \cdot \frac{\lambda_3(\lambda_3-1) y_i^{(\lambda_3-2)} - \lambda_4(\lambda_4-1)(1-y_i)^{(\lambda_4-2)}}{\left(\lambda_3 y_i^{(\lambda_3-1)} + \lambda_4(1-y_i)^{(\lambda_4-1)}\right)^2}$$

$$\frac{\partial \log L_0}{\partial \lambda_2} = \frac{n}{\lambda_2} + \sum_{i=1}^{n} \frac{1}{\lambda_2} \cdot \frac{\left(\lambda_3(\lambda_3-1) y_i^{(\lambda_3-2)} - \lambda_4(\lambda_4-1)(1-y_i)^{(\lambda_4-2)}\right)\left(y_i^{\lambda_3} - (1-y_i)^{\lambda_4}\right)}{\left(\lambda_3 y_i^{(\lambda_3-1)} + \lambda_4(1-y_i)^{(\lambda_4-1)}\right)^2}$$

$$\frac{\partial \log L_0}{\partial \lambda_3} = -\sum_{i=1}^{n} \frac{\left(\lambda_3(\lambda_3-1) y_i^{(\lambda_3-2)} - \lambda_4(\lambda_4-1)(1-y_i)^{(\lambda_4-2)}\right)\left(\log y_i \cdot y_i^{\lambda 3}\right)}{\left(\lambda_3 y_i^{(\lambda_3-1)} + \lambda_4(1-y_i)^{(\lambda_4-1)}\right)^2}$$

$$\frac{\partial \log L_0}{\partial \lambda_4} = -\sum_{i=1}^{n} \frac{\left(\lambda_3(\lambda_3-1) y_i^{(\lambda_3-2)} - \lambda_4(\lambda_4-1)(1-y_i)^{(\lambda_4-2)}\right)\left(\log(1-y_i) \cdot (1-y_i)^{\lambda_4}\right)}{\left(\lambda_3 y_i^{(\lambda_3-1)} + \lambda_4(1-y_i)^{(\lambda_4-1)}\right)^2}$$

We set all the above nonlinear equations to be 0 and obtain the estimates of all the parameters under H_0 and then

$$\mathrm{SIC}(n) = -2\left[n \cdot \log \hat{\lambda}_2 - \sum_{i=1}^{n} \log\left(\hat{\lambda}_3 y_i^{\hat{\lambda}_3-1} + \hat{\lambda}_4 (1-y_i)^{\hat{\lambda}_4-1}\right)\right] + 4 \cdot \log n, \tag{9}$$

where y_i's are obtained from Eq. (8). Under H_1, the likelihood function is

$$\begin{aligned} L_1 &= \prod_{i=1}^{k} f(x_i; \boldsymbol{\theta_1}) \cdot \prod_{i=k+1}^{n} f(x_i; \boldsymbol{\theta_2}) \\ &= \prod_{i=1}^{n} \frac{\lambda_2^{(1)}}{\lambda_3^{(1)} y_i^{(\lambda_3^{(1)}-1)} + \lambda_4^{(1)} (1-y_i)^{(\lambda_4^{(1)}-1)}} \cdot \prod_{i=k+1}^{n} \frac{\lambda_2^{(n)}}{\lambda_3^{(n)} y_i^{(\lambda_3^{(n)}-1)} + \lambda_4^{(n)} (1-y_i)^{(\lambda_4^{(n)}-1)}}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} x_i &= Q(y_i) = \lambda_1^{(1)} + \frac{y_i^{\lambda_3^{(1)}} - (1-y_i)^{\lambda_4^{(1)}}}{\lambda_2^{(1)}}, \quad i = 1, 2, \ldots, k; \\ x_i &= Q(y_i) = \lambda_1^{(n)} + \frac{y_i^{\lambda_3^{(n)}} - (1-y_i)^{\lambda_4^{(n)}}}{\lambda_2^{(n)}}, \quad i = k+1, \ldots, n, \end{aligned} \tag{11}$$

$\boldsymbol{\theta_1} = (\lambda_1^{(1)}, \lambda_2^{(1)}, \lambda_3^{(1)}, \lambda_4^{(1)})$ and $\boldsymbol{\theta_2} = (\lambda_1^{(n)}, \lambda_2^{(n)}, \lambda_3^{(n)}, \lambda_4^{(n)})$, As before, we obtain the following nonlinear equations based on $\log L_1$:

$$\frac{\partial \log L_1}{\partial \lambda_1^{(1)}} = -\left[\lambda_2^{(1)} \sum_{i=1}^{k} \frac{\lambda_3^{(1)}(\lambda_3^{(1)} - 1)y_i^{(\lambda_3^{(1)}-2)} - \lambda_4^{(1)}(\lambda_4^{(1)} - 1)(1 - y_i)^{(\lambda_4^{(1)}-2)}}{\left(\lambda_3^{(1)} y_i^{(\lambda_3^{(1)}-1)} + \lambda_4^{(1)}(1 - y_i)^{(\lambda_4^{(1)}-1)}\right)^2}\right]$$

$$\frac{\partial \log L_1}{\partial \lambda_1^{(n)}} = -\left[\lambda_2^{(n)} \sum_{i=k+1}^{n} \frac{\lambda_3^{(n)}(\lambda_3^{n} - 1)y_i^{(\lambda_3^{(n)}-2)} - \lambda_4^{(n)}(\lambda_4^{(n)} - 1)(1 - y_i)^{(\lambda_4^{(n)}-2)}}{\left(\lambda_3^{(n)} y_i^{(\lambda_3^{n}-1)} + \lambda_4^{(n)}(1 - y_i)^{(\lambda_4^{(n)}-1)}\right)^2}\right]$$

$$\frac{\partial \log L_1}{\partial \lambda_2^{(1)}} = \frac{k}{\lambda_2^{(1)}} + \sum_{i=1}^{n} \frac{1}{\lambda_2^{(1)}} \cdot \frac{\begin{array}{c}\left(\lambda_3^{(1)}(\lambda_3^{(1)} - 1)y_i^{(\lambda_3^{(1)}-2)} - \lambda_4^{(1)}(\lambda_4^{(1)} - 1)(1 - y_i)^{(\lambda_4^{(1)}-2)}\right)\\ \left(y_i^{\lambda_3^{(1)}} - (1 - y_i)^{\lambda_4^{(1)}}\right)\end{array}}{\left(\lambda_3^{(1)} y_i^{(\lambda_3^{(1)}-1)} + \lambda_4^{(1)}(1 - y_i)^{(\lambda_4^{(1)}-1)}\right)^2}$$

$$\frac{\partial \log L_1}{\partial \lambda_2^{(n)}} = \frac{n-k}{\lambda_2^{(n)}} + \sum_{i=k+1}^{n} \frac{1}{\lambda_2^{(n)}} \cdot \frac{\begin{array}{c}\left(\lambda_3^{(n)}(\lambda_3^{(n)} - 1)y_i^{(\lambda_3^{(n)}-2)} - \lambda_4^{(n)}(\lambda_4^{(n)} - 1)(1 - y_i)^{(\lambda_4^{(n)}-2)}\right)\\ \left(y_i^{\lambda_3^{(n)}} - (1 - y_i)^{\lambda_4^{(n)}}\right)\end{array}}{\left(\lambda_3^{(n)} y_i^{(\lambda_3^{(n)}-1)} + \lambda_4^{(n)}(1 - y_i)^{(\lambda_4^{(n)}-1)}\right)^2}$$

$$\frac{\partial \log L_1}{\partial \lambda_3^{(1)}} = -\sum_{i=1}^{k} \frac{\left(\lambda_3^{(1)}(\lambda_3^{(1)} - 1)y_i^{(\lambda_3^{(1)}-2)} - \lambda_4^{(1)}(\lambda_4^{(1)} - 1)(1 - y_i)^{(\lambda_4^{(1)}-2)}\right)\left(\log y_i \cdot y_i^{\lambda_3^{(1)}}\right)}{\left(\lambda_3^{(1)} y_i^{(\lambda_3^{(1)}-1)} + \lambda_4^{(1)}(1 - y_i)^{(\lambda_4^{(1)}-1)}\right)^2}$$

$$\frac{\partial \log L_1}{\partial \lambda_3^{(n)}} = -\sum_{i=k+1}^{n} \frac{\left(\lambda_3^{(n)}(\lambda_3^{(n)} - 1)y_i^{(\lambda_3^{(n)}-2)} - \lambda_4^{(n)}(\lambda_4^{(n)} - 1)(1 - y_i)^{(\lambda_4^{(n)}-2)}\right)\left(\log y_i \cdot y_i^{\lambda_3^{(n)}}\right)}{\left(\lambda_3^{(n)} y_i^{(\lambda_3^{(n)}-1)} + \lambda_4^{(n)}(1 - y_i)^{(\lambda_4^{(n)}-1)}\right)^2}$$

$$\frac{\partial \log L_1}{\partial \lambda_4^{(1)}} = -\sum_{i=1}^{n} \frac{\begin{array}{c}\left(\lambda_3^{(1)}(\lambda_3^{(1)} - 1)y_i^{(\lambda_3^{(1)}-2)} - \lambda_4^{(1)}(\lambda_4^{(1)} - 1)(1 - y_i)^{(\lambda_4^{(1)}-2)}\right)\\ \left(\log(1 - y_i) \cdot (1 - y_i)^{\lambda 4^{(1)}}\right)\end{array}}{\left(\lambda_3^{(1)} y_i^{(\lambda_3^{(1)}-1)} + \lambda_4^{(1)}(1 - y_i)^{(\lambda_4^{(1)}-1)}\right)^2}$$

$$\frac{\partial \log L_1}{\partial \lambda_4^{(n)}} = -\sum_{i=k+1}^{n} \frac{\begin{array}{c}\left(\lambda_3^{(n)}(\lambda_3^{(n)} - 1)y_i^{(\lambda_3^{(n)}-2)} - \lambda_4^{(n)}(\lambda_4^{(n)} - 1)(1 - y_i)^{(\lambda_4^{(n)}-2)}\right)\\ \left(\log(1 - y_i) \cdot (1 - y_i)^{\lambda 4^{(n)}}\right)\end{array}}{\left(\lambda_3^{(n)} y_i^{(\lambda_3^{(n)}-1)} + \lambda_4^{(n)}(1 - y_i)^{(\lambda_4^{(n)}-1)}\right)^2}.$$

We obtain the MLEs of all eight parameters $\hat{\lambda}_i^{(1)}, \hat{\lambda}_i^{(n)}, i = 1, 2, 3, 4$ by setting the above equations to 0. Then

$$\begin{aligned}\mathrm{SIC}(k) = -2\Bigg[& k \cdot \log \hat{\lambda}_2^{(1)} + (n-k) \cdot \log \hat{\lambda}_2^{(n)} - \sum_{i=1}^{k} \log\left(\hat{\lambda}_3^{(1)} y_i^{\hat{\lambda}_3^{(1)}-1} + \hat{\lambda}_4^{(1)}(1 - y_i)^{\hat{\lambda}_4^{(1)}-1}\right) \\ & - \sum_{i=k+1}^{n} \log\left(\hat{\lambda}_3^{(n)} y_i^{\hat{\lambda}_3^{(n)}-1} + \hat{\lambda}_4^{(n)}(1 - y_i)^{\hat{\lambda}_4^{(n)}-1}\right)\Bigg] + 8 \cdot \log n, \end{aligned} \tag{12}$$

where y_i’s are obtained from Eq. (11).

To fit a data set with a single GLD, we use the R package GLDEX developed by Su (2007) to obtain the estimates for λ_i's. Then we calculate the values of SIC(k) for $2 \leq k \leq n-2$.

4. Asymptotic Null Distribution

When we use SIC to detect the change points, it may happen that SIC values are very close which may due to the random noises from the data. Therefore, it may not indicate the true change points. To make the conclusion more statistically convincing, we follow Chen and Gupta (2000) to consider the following test statistic:

$$\Delta_n = \min_{2\leq k\leq n-2}[\mathrm{SIC}(k) - \mathrm{SIC}(n)]$$

and study its asymptotic distribution.

We accept H_0 if $\mathrm{SIC}(n) < \min_{2\leq k\leq n-2}\mathrm{SIC}(k) + c_\alpha$, instead of $\mathrm{SIC}(n) < \min_{2\leq k\leq n-2}\mathrm{SIC}(k)$, where c_α is determined by $1-\alpha = P(\mathrm{SIC}(n) < \min \mathrm{SIC}(k) + c_\alpha \mid H_0)$. We define

$$Z_n = \max_{2\leq k\leq n-2}\{-2\log(L_0/L_1)\}.$$

Csörgő and Horváth (1997) showed that

$$\lim_{n\to\infty} P\{A(\log n)Z_n^{1/2} \leq x + D_4(\log n)\} = \exp(-2e^{-t}) \tag{13}$$

where $A(\log n) = (2\log\log n)^{1/2}$ and $D_4(\log n) = 2\log\log n + 2\log\log\log n$. Since

$$\begin{aligned}\Delta_n &= \min_{2\leq k\leq n-2}[\mathrm{SIC}(k) - \mathrm{SIC}(n)]\\ &= -\max_{2\leq k\leq n-2}[\mathrm{SIC}(n) - \mathrm{SIC}(k)]\\ &= -\max_{2\leq k\leq n-2}[-2\log L_0 + 4\log n - (-2\log L_1 + 8\log n)]\\ &= -\max_{2\leq k\leq n-2}[-2(\log L_0 - \log L_1) - 4\log n]\\ &= -Z_n + 4\log n;\end{aligned}$$

therefore, we have

$$Z_n = (4\log n - \Delta_n).$$

With the result in (13), we obtain that

$$\lim_{n\to\infty} P\{A(\log n)(4\log n - \Delta_n)^{1/2} - D_4(\log n) \leq x\} = \exp(-2e^{-t}). \tag{14}$$

The approximated c_α values can be determined as follows:

$$\begin{aligned}1-\alpha &= P\Big(\mathrm{SIC}(n) < \min_{2\leq k\leq n-2}\mathrm{SIC}(k) + c_\alpha \mid H_0\Big) = P\Big(\mathrm{SIC}(n) - \min_{2\leq k\leq n-2}\mathrm{SIC}(k) < c_\alpha \mid H_0\Big)\\ &= P\Big(-\max_{2\leq k\leq n-2}(\mathrm{SIC}(n) - \mathrm{SIC}(k)) \geq c_\alpha\Big) = P(\Delta_n \geq c_\alpha)\end{aligned}$$

$$
\begin{aligned}
&= P(4\log n - Z_n \geq c_\alpha) = p(0 \leq Z_n^{1/2} \leq (4\log n + c_\alpha)^{1/2}) \\
&= P(-D_4(\log n) \leq A(\log n)Z_n^{1/2} - D_4(\log n) \\
&\leq A(\log n)(4\log n + c_\alpha)^{1/2} - D_4(\log n)) \\
&\cong \exp\{-2\exp[D_4(\log n) - A(\log n)(4\log n + c_\alpha)^{1/2}]\} - \exp\{-2\exp[D_4(\log n)]\}.
\end{aligned}
$$

Then we can obtain the approximate values of c_α at different levels α by solving the above equation:

$$
c_\alpha \cong \left[\frac{D_4(\log n)}{A(\log n)} - \frac{1}{A(\log n)}\log\log\left[1 - \alpha + \exp(-2e^{D_4(\log n)})\right]^{-\frac{1}{2}}\right]^2 - 4\log n. \quad (15)
$$

For different significance levels $\alpha = 0.010, 0.025, 0.05$, and 0.100, and various sample sizes $n = 7, 8, \ldots, 200$, we compute the critical values for an adjusted SIC procedure proposed above based on (15) (see Table 1 in the Appendix).

5. Simulation Study

In this section, we conduct simulation studies on the GLD change point model for moderate sample sizes 30, 40, 50, and 60. We choose the true location of the change point at the front (the $\frac{[n]}{3}$th position), the middle (the $\frac{n}{2}$th position), and the end (the $\frac{[2n]}{3}$th position). Before the change point, the observations follow a GLD(0, 1, 0.19, 0.19), and after the change point, the observations will follow GLD(0.4, 1.4, 0.59, 0.59), GLD(0.5, 1.5, 0.69, 0.69), GLD(0.6, 1.6, 0.79, 0.79), respectively. The powers for the testing procedure using the GLD and the normal distribution are listed as follows.

	$n = 30$		$n = 40$		$n = 50$		$n = 60$	
	GLD	Normal	GLD	Normal	GLD	Normal	GLD	Normal
Front	0.823	0.782	0.976	0.878	0.990	0.892	1.000	0.924
Middle	0.934	0.886	0.972	0.916	0.966	0.931	1.000	0.978
End	0.852	0.778	0.921	0.884	0.973	0.914	0.976	0.938
Front	0.925	0.845	0.982	0.864	0.992	0.860	1.000	0.893
Middle	0.984	0.902	0.996	0.907	0.999	0.922	1.000	0.958
End	0.934	0.817	0.990	0.897	1.000	0.926	1.000	0.932
Front	0.966	0.756	0.994	0.733	1.000	0.847	1.000	0.831
Middle	0.999	0.812	1.000	0.865	1.000	0.921	1.000	0.955
End	1.000	0.845	1.000	0.833	1.000	0.913	1.000	0.929

The above simulation results indicate that the proposed method has a high power to identify the locations of change points in the middle or at the end of the data, and that the power increases as the sample size increases. When the sample size is 30, and the difference of the parameters between two GLDs before and after a change point is small, the power of detecting the change point located at the beginning or at the end the data set is around 82 or 85%. However, as the sample size increases, the power increases quickly (from 0.82 to 1.00, and from 0.85 to 0.97). We also can observe that the power increases and gets closer to 1 as the difference of the parameters between GLDs before and after a change point increases. From the above table, we also observe that the power of the testing procedure with the GLD change point model is higher than the one with the normal change point model, especially when the sample size is small ($n = 30$, 40). The reason is that simulated data is skewed not symmetric. Therefore, the model with the normal distribution can not handle it well. However, the model with the GLD can handle such a situation quite well since it has four parameters to control the location, scale, skewness and kurtosis at the same time. The simulation results indicate that the GLD family is more flexible than the normal distribution family, especially when fitting a non symmetric data.

6. Application to Real Data

Snijders et al. (2001) performed array comparative genomic hybridization (CGH) experiment on 15 fibroblast cell lines based on the normalized averages of the log $_2T_i/R_i$ along positions on each chromosome. The data is available at the fibroblast cell line data website (http://www.nature.com/ng/journal/v29/n3/full/ng754.html). In Table 1 of Snijders et al. (2001), they indicated the verified copy number of variations for all 23 chromosomes for each cell line by karyotyping. In Chen and Wang (2008), they conducted a chromosome wide search for all 23 chromosomes within each of the 9 chosen cell lines in order to identify the copy number changes in a mean and variance change point model (MVCM). We now choose one of these 15 cell lines, GM01750, and use the GLD change point model to conduct a chromosome wide search. Among all 23 chromosomes of GM01750, on chromosome 9, we found that

$$\text{SIC}(23) = -211.60829 = \min_{2\leq k\leq 103} \text{SIC}(k) < \text{SIC}(105) = -63.71836.$$

If we use c_α values at different significance levels with $n = 105$ in Table 1 (see Appendix), we still find that

$$\text{SIC}(23) = -211.60829 + c_\alpha < \text{SIC}(105) = -63.71836$$

which indicates that the position 24 is a change point on the chromosome 9 of the fibroblast cell line GM01750. The result matches with the copy number variations identified through the spectral karyotyping (Table 1, Snijders et al., 2001). Figure 1 shows the plot of SIC values for $2 \leq k \leq 103$, and the plot of genomic positions of chromosome 9.

Chromosome 14 is another chromosome of GM01750 in which we found a change point since

$$\text{SIC}(10) = -146.89965 = \min_{2\leq k\leq 73} \text{SIC}(k) < \text{SIC}(75) = -73.80452.$$

If we use c_α values at different significance levels with $n = 75$ in Table 1 (see Appendix), we still find that

$$\text{SIC}(10) = -146.89965 + c_\alpha < \text{SIC}(75) = -73.80452$$

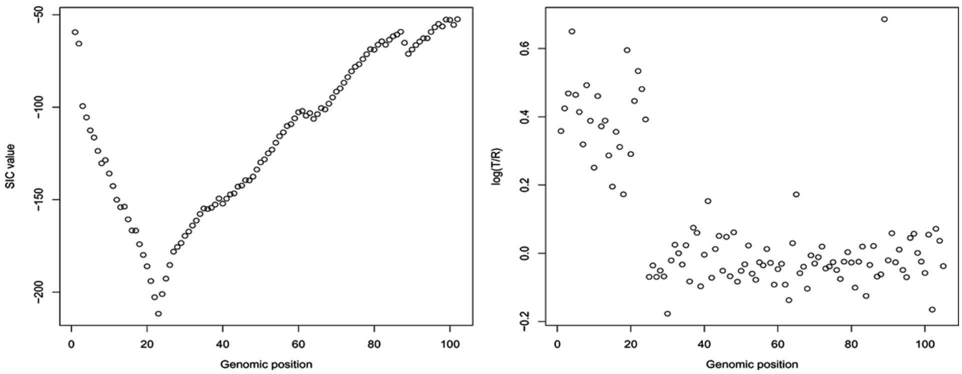


Figure 1. Left: SIC values for chromosome 9 of the fibroblast cell line with genomic position from 2 to 103; Right: Scatterplot for chromosome 9 of the fibroblast cell line with genomic position from 1 to 105.

which indicates that the position 11 is a change point on the chromosome 14 of the fibroblast cell line GM01750. The result matches with the copy number variations identified through the spectral karyotyping (Table 1, Snijders et al., 2001). Figure 2 shows the plot of SIC values for $2 \leq k \leq 73$, and the plot of genomic positions of chromosome 14. For both chromosomes, we use the binary method to keep testing the subsequences after we locate the first change points, and find no more change points.

7. Discussion

In this article, we study the change point problem for the generalized lambda distributions (GLDs). With the combination of the binary segmentation method and information approach (SIC), we detect possible change points in a data by fitting a single GLD. Simulation results in Sec. 5 indicate that the GLD change point model is more powerful than the normal change point model on detecting change

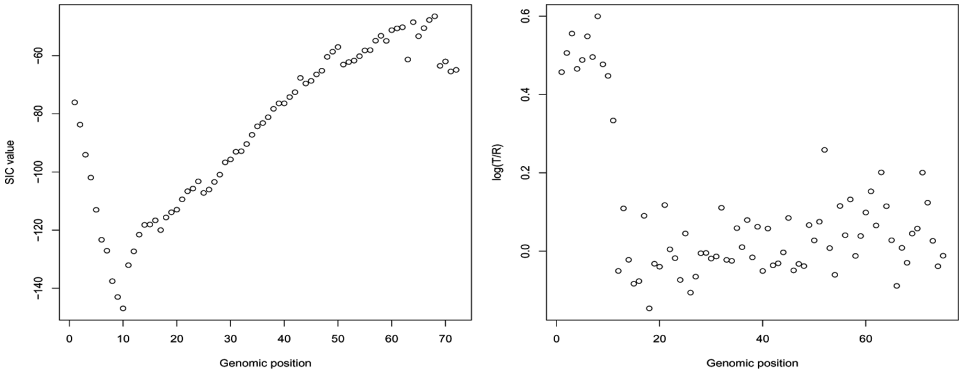


Figure 2. Left: SIC values for chromosome 14 of the fibroblast cell line with genomic position from 2 to 73; Right: Scatterplot for chromosome 14 of the fibroblast cell line with genomic position from 1 to 75.

points for a non-symmetric data, thus, it shows the GLD family is more flexible and broad. To make our conclusions statistically convincing, we also derive the critical values at different significant levels for the testing procedure. The likelihood ratio test (LRT) is an alternative to test the hypotheses of change points. However, due to the complexity of the density function of a GLD, the test statistic does not have a simple form as for the other distributions, such as the normal distribution or the exponential distribution. One way to obtain its approximate null distribution is through MCMC, which we will explore in our future work. To obtain MLEs of the parameters, Eqs. (5)–(11) may not have solutions. In such a situation, we use Nelder and Mead (1965) simplex algorithm to find the values which will maximize the likelihood function, instead of solving those equations directly. The R package GLDEX developed by Su (2007) provides such an alternative.

Appendix

Table 1
Approximate critical values of SIC

n/α	0.100	0.050	0.025	0.010
7	17.715	25.503	34.148	46.968
8	17.181	24.969	33.614	46.433
9	16.710	24.498	33.143	45.962
10	16.288	24.076	32.722	45.541
11	15.907	23.695	32.341	45.160
12	15.559	23.347	31.992	44.812
13	15.239	23.027	31.672	44.491
14	14.942	22.730	31.376	44.195
15	14.666	22.454	31.100	43.919
16	14.408	22.196	30.842	43.661
17	14.166	21.954	30.599	43.418
18	13.937	21.725	30.371	43.190
19	13.721	21.509	30.154	42.973
20	13.516	21.304	29.949	42.768
21	13.321	21.208	29.754	42.573
22	13.134	20.922	29.568	42.387
23	12.957	20.745	29.390	42.209
24	12.786	20.574	29.220	42.039
25	12.623	20.411	29.056	41.876
26	12.466	20.254	28.900	41.719
27	12.315	20.103	28.749	41.568
28	12.170	19.958	28.603	41.422
29	12.029	19.817	28.463	41.282
30	11.894	19.682	28.327	41.146
35	11.277	19.065	27.711	40.530
40	10.743	18.531	27.177	39.996
45	10.272	18.060	26.705	39.524

(continued)

Table 1
Continued

n/α	0.100	0.050	0.025	0.010
50	9.851	17.638	26.284	39.103
55	9.469	12.257	25.903	38.722
60	9.121	16.910	25.555	38.374
65	8.801	16.589	25.235	38.054
70	8.505	16.293	24.938	37.757
75	8.229	16.017	24.662	37.481
80	7.971	15.785	24.404	37.223
85	7.728	15.516	24.161	36.981
90	7.499	15.287	23.933	36.752
100	7.078	14.866	23.511	36.330
105	6.883	14.671	23.316	36.135
120	6.349	14.137	22.782	35.601
140	5.732	13.520	22.166	34.985
160	5.198	12.986	21.631	34.450
180	4.729	12.515	21.160	33.979
200	4.305	12.093	20.739	33.558

Acknowledgments

The authors wish to thank the two referees for their helpful comments, which helped the article's clarity and crispness. This work was partly done when A.K. Gupta was on FIL at Department of Biostatistics, University of Michigan, Ann Arbor, Michigan.

References

Barry, D., Hartigan, J. A. (1993). A Bayesian analysis for change point problem. *Journal of American Statistical Association* 88:309–319.

Chen, J., Gupta, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of American Statistical Association* 92:739–747.

Chen, J., Gupta, A. K. (2000). *Parametric Statistical Change Point Analysis*. Boston: Birkhäuser.

Chen, J., Wang, Y. P. (2008). A Statistical Change Point Model Approach for the Detection of DNA Copy Number Variations in Array CGH Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 25. IEEE computer Society Digital Library. ⟨http://doi.ieeecomputersociety.org/10.1109/TCBB.2008.129⟩.

Chernoff, H., Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics* 35:999–1018.

Csörgő, M., Horváth, L. (1997). *Limit Theorems in Change-Point Analysis.* New York: John Wiley & Sons.

Gardner, L. A. (1969). On detecting change in the mean of normal variates. *Annals of Mathematical Statistics* 40:116–126.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics Theory* 19(8): 2745–2756.

Hawkins, D. M. (1992). Detecting shifts in functions of multivaraite location and covariance parameters. *Journal of Statistical Planning and Inference* 33:233–244.

Hsu, D. A. (1977). Tests for variance shifts at an unknown time point. *Applied Statistics* 26:179–184.

Hsu, D. A. (1979). Detecting shifts of parameters in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association* 74:31–40.

Inclán, C. (1993). Detection of multiple changes of variance using posterior odds. *Journal of Business and Economic Statistics* 11:189–300.

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

Kim, H. J., Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika* 76:409–423.

Krishnaiah, P. R., Miao, B. Q. (1988). Review about estimation of change points. In: Krishnaiah, P. R., Rao, C. R., eds. *Handbook of Statistics*, Vol. 7. Amsterdam: Elsevier, pp. 375–402.

Loschi, R. H., Cruz, F. R. B., Arellano-Valle, R. B. (2005). Multiple change point analysis for the regular exponential family using the product partition model. *Journal of Data Science* 3:305–330.

Nelder, J. A., Mead, R. (1965). A simplex method for function minimization. *Computer Journal* 7:308–313.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. *Philisophical Transactions of the Royal Society of London A* 185:71–110.

Ramberg, J. S., Schmeiser, B. W. (1972). An approximate method for generating symmetric random variables. *Communications of the ACM* 15:987–990.

Ramberg, J. S., Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM* 17:78–82.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics* 21:201–204.

Sen, A. K., Srivastava, M. S. (1975). On tests for detecting change in mean. *The Annals of Statistics* 3:98–108.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D., Alberston, D. G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* 29:263–264.

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

Tukey, J. W. (1960). The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts. Technical Reports, 36, Statistical Techniques Research Group, Princeton University.

Wichern, D. W., Miller, R. B., Hsu, D. A. (1976). Changes of variance in first order autoaggressive time series models – with an application. *Applied Statistics* 25:248–356.

Worsley, K. J. (1979). On the likelihood ratio tests for a shift in location of normal populations. *Journal of the American Statistical Association* 74:365–367.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* 73:91–104.

Vostrikova, L. J. (1981). Detecting 'disorder' in multidimensional random process. *Soviet Mathematics Doklady* 24:55–59.