

Understanding and Addressing the Unbounded “Likelihood” Problem

Shiyao Liu, Huaqing Wu, William Q. Meeker

Department of Statistics

Iowa State University

August 2, 2013

Abstract

The joint probability density function, evaluated at the observed data, is commonly used as the likelihood function to compute maximum likelihood estimates. For some models, however, there exist paths in the parameter space along which this density-approximation likelihood goes to infinity and maximum likelihood estimation breaks down. In applications, all observed data are discrete due to the round-off or grouping error of measurements. The “correct likelihood” based on interval censoring can eliminate the problem of an unbounded likelihood. This paper categorizes the models leading to unbounded likelihoods into three groups and illustrates the density breakdown with specific examples. We also study the effect of the round-off error on estimation, and give a sufficient condition for the joint density to provide the same maximum likelihood estimate as the correct likelihood, as the round-off error goes to zero.

Key words: Density approximation; Interval censoring; Maximum likelihood; Round-off error; Unbounded likelihood.

1 Introduction

1.1 Background

Because of inherent limitations of measuring instruments, all continuous numerical data are subject to the round-off or grouping error of measurements. This has been described, for example, by Kempthorne (1966), Barnard (1967), Kempthorne and Folks (1971), Giesbrecht and Kempthorne (1976), Cheng and Iles (1987), and Vardeman and Lee (2005).

For the sake of discussion, suppose that data come from an accurate source with no bias or measurement error (the usual tacit assumption in elementary statistical analysis). An example would be a properly calibrated digital voltmeter. Rounding is controlled by the choice of how many digits to display or record, after the decimal place. If the voltmeter rounds to the nearest integer, there would be a rounding uncertainty of plus/minus $\Delta = 0.5$; if the voltmeter rounds to the nearest 0.1, there would be a rounding uncertainty of plus/minus $\Delta = 0.05$, and so on. Other rounding conventions could be followed, but would not have any effect on the main points or conclusions to be presented in this paper.

For convenience, such discrete observations are often modeled on a continuous scale. Usually, when the round-off error is small, the likelihood for a sample of independent observations is defined as the product of the probability densities evaluated at each of the “exact” observations. For some models, however, such a likelihood may be unbounded along certain paths in the parameter space, causing numerical and statistical problems in maximum likelihood (ML) estimation. As has been suggested in the references above, using the *correct likelihood* based on small intervals (e.g., implied by the data’s precision) instead of the density approximation will eliminate the problem of an unbounded likelihood. Practitioners should know about the potential problems of an unbounded likelihood and how to use the correct likelihood to avoid the problems. The purpose of this paper is to review and consolidate previous results concerning this problem, to provide a classification of models that lead to an unbounded “likelihood,” and to present related theoretical results.

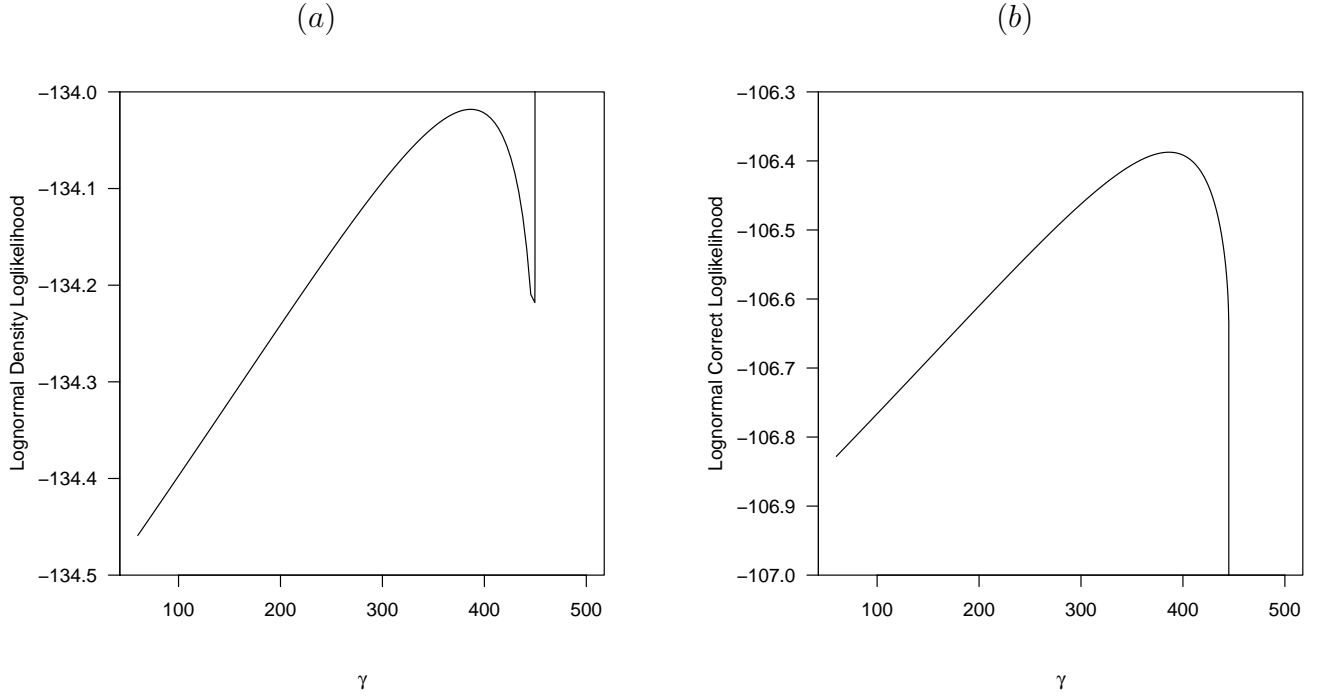


Figure 1: Three-parameter lognormal profile log-likelihood plots of the threshold parameter γ for the diesel generator fan data using (a) the unbounded density-approximation likelihood L and (b) the correct likelihood \mathcal{L} (with the round-off error $\Delta = 5$).

1.2 An Illustrative Example

Example 11.17 of Meeker and Escobar (1998) illustrates ML estimation for the three-parameter lognormal distribution using the diesel generator fan data given in Nelson (1982, page 133). The likelihood function based on the usual density approximation for exact failures at time t_i , $i = 1, \dots, n$ has the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n L_i(\boldsymbol{\theta}; t_i) = \prod_{i=1}^n f(t_i; \boldsymbol{\theta}), \quad (1)$$

where

$$f(t_i; \boldsymbol{\theta}) = \frac{1}{\sigma(t_i - \gamma)} \phi_{\text{nor}} \left[\frac{\log(t_i - \gamma) - \mu}{\sigma} \right] I_{(t_i > \gamma)} \quad (2)$$

is the probability density function (pdf) of the three-parameter lognormal distribution and $\boldsymbol{\theta} = (\mu, \sigma, \gamma)'$. Here ϕ_{nor} is the pdf for the standard normal distribution and $\exp(\mu)$, σ and

γ are the scale, shape and threshold parameters, respectively. As the threshold parameter γ approaches the smallest observation $t_{(1)}$, the profile log-likelihood for γ in Figure 1 (a) increases without bound (i.e., $L(\boldsymbol{\theta}) \rightarrow \infty$), indicating the breakdown in the density approximation. For the diesel generator fan data, there is a local maximum that corresponds to the maximum of the correct likelihood. For some data sets, the local maximum is dominated by the unbounded behavior.

1.3 A Simple Remedy

The unboundedness of the likelihood can cause computational difficulties or convergence to a nonsense maximum of the density-approximation likelihood. As suggested in the references in Section 1.1, using the “correct likelihood” that acknowledges rounding in the data and is based on interval censoring (instead of the density-approximation likelihood) will eliminate the problem of an unbounded likelihood. Because probabilities can not be larger than 1, the correct likelihood will always be bounded. The correct likelihood can be expressed as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}; t_i) = \prod_{i=1}^n \frac{1}{2\Delta_i} \int_{t_i - \Delta_i}^{t_i + \Delta_i} f(x; \boldsymbol{\theta}) dx \\ &= \prod_{i=1}^n \frac{1}{2\Delta_i} [F(t_i + \Delta_i; \boldsymbol{\theta}) - F(t_i - \Delta_i; \boldsymbol{\theta})],\end{aligned}\tag{3}$$

where, for the example in Section 1.2,

$$F(t_i; \boldsymbol{\theta}) = \Phi_{\text{nor}} \left[\frac{\log(t_i - \gamma) - \mu}{\sigma} \right] \mathbf{I}_{(t_i > \gamma)}$$

is the three-parameter lognormal cumulative distribution function (cdf). Here Φ_{nor} is the cdf for the standard normal distribution.

The values of Δ_i reflect the round-off error in the data and may depend on the magnitude of the observations. For the diesel generator fan data, because the life times were recorded to a precision of ± 5 hours, we choose $\Delta_i = 5$ for all of the t_i values. Figure 1 (b) shows that, with the correct likelihood, the profile plot is well behaved with a clear maximum at a value of γ that is a little less than 400.

1.4 R. A. Fisher’s Definition of Likelihood

Fisher (e.g., 1912, page 157) suggests that a likelihood defined by a product of densities should be proportional to the probability of the data (which we now know is often, but not always true). In particular, he says “... then P' [the joint density] is proportional to the chance of a given set of observations occurring.” Fisher (1922, page 327) points out that “Likelihood [expressed as a joint probability density] also differs from probability in that it is not a differential element, and is incapable of being integrated: it is assigned to a particular point of the range of variation, not to a particular element of it.”

1.5 Related Literature

Cheng and Iles (1987) summarize several alternative methods of estimation that have been proposed to remedy the unbounded likelihood problem. Cheng and Amin (1983) suggest the maximum product of spacings (MPS) method. This method can be applied to any univariate distribution. Wong and Li (2006) use the MPS method to estimate the parameters of the maximum generalized extreme value (GEV) distribution and the generalized Pareto distribution (GPD), both of which have unbounded density-approximation likelihood functions. Cheng and Traylor (1995) point out the drawbacks of the MPS method owing to the occurrence of the tied observations and numerical effects involved in ordering the cdf when there are explanatory variables in the model. Harter and Moore (1965) suggest that one can use the smallest observation to estimate the threshold parameter and then estimate the other two parameters using the remaining observations. This method has been further studied by Smith and Weissman (1985). Although the smallest observation is the ML estimator of the threshold parameter, under this method, the ML estimators of the other parameters are no longer consistent. Kempthorne (1966) and Barnard (1967) suggest a method that is similar to the interval-censoring approach; their method groups the observations into non-overlapping cells, implying a multinomial distribution in which the cell probabilities depend on the unknown parameters.

Cheng and Traylor (1995) describe the unbounded likelihood problem as one of the four types of non-regular maximum likelihood problems with specific examples including the

three-parameter Weibull distribution and discrete mixture models. Cheng and Amin (1983) point out that in the three-parameter lognormal, Weibull and gamma distributions, there exist paths in the parameter space where as the threshold parameter tends to the smallest observation, the density-approximation likelihood function approaches infinity. Giesbrecht and Kempthorne (1976) show that the unbounded likelihood problem of the three-parameter lognormal distribution can be overcome by using the correct likelihood instead of the density approximation. Atkinson, Pericchi, and Smith (1991) apply the grouped-data likelihood approach to the shifted power transformation model of Box and Cox (1964). Kulldorff (1957, 1961) argues that ML estimators based on the correct likelihood for grouped data are consistent and asymptotically efficient. Other examples of unbounded density-approximation likelihood functions are given in Section 2.

1.6 Overview

The remainder of this paper is organized as follows. Section 2 divides the models leading to unbounded likelihoods into three categories and for each category, illustrates the density-approximation breakdown with specific examples frequently encountered in practice. Section 3, using the minimum GEV distribution and the mixture of two univariate normal distributions (both of which have unbounded likelihood functions) as examples, studies the effect that different amounts of the round-off error have on estimation. Section 4 describes the equicontinuity condition, which is a sufficient condition for the product of densities to provide the same maximum likelihood estimate as the correct likelihood, as the round-off error goes to zero. Section 5 provides some conclusions.

2 A Classification of Unbounded “Likelihoods”

In this section, we divide the models that have an unbounded density-approximation likelihood into three categories.

- Continuous univariate distributions with three or four parameters, including a threshold parameter.

- Discrete mixture models of continuous distributions for which at least one component has both a location and a scale parameter.
- Minimum-type (and maximum-type) models for which at least one of the marginal distributions has both a location and a scale parameter.

The classification we provide includes all of the unbounded likelihood situations that we have observed or found in the literature. Our classification, however, may not be exhaustive.

2.1 Continuous Univariate Distributions with Three or Four Parameters, Including a Threshold Parameter

For n independent and identically distributed (iid) observations x_1, x_2, \dots, x_n from a certain distribution with a threshold parameter that shifts the distribution by an amount γ , the pdf is $f(x) > 0$ for all $x > \gamma$ and $f(x) = 0$ for $x \leq \gamma$. Generally, there exist paths in the parameter space where the “likelihood” function grows without bound as the threshold parameter tends to the smallest observation. For example, in the log-location-scale distributions (e.g., Weibull, Fréchet, loglogistic and lognormal) with a threshold parameter, the pdf is

$$\frac{1}{\sigma(x - \gamma)} f_0 \left(\frac{\log(x - \gamma) - \mu}{\sigma} \right) \mathbf{I}_{(x > \gamma)}, \quad (4)$$

where $f_0(x) > 0$ for all x is the pdf of a location-scale distribution. Here $\exp(\mu)$ is a scale parameter, $\sigma > 0$ is a shape parameter, and γ is a threshold parameter. The density-approximation likelihood function $L(\mu, \sigma, \gamma) \rightarrow \infty$ as $\mu = \log(x_{(1)} - \gamma)$, $\sigma^2 = \mu^2$, and $\gamma \rightarrow x_{(1)}$ (with $\gamma < x_{(1)}$).

The simple example in Section 1.2 using the three-parameter lognormal distribution is an example in this category. The profile plot in Figure 1 (a) indicates the breakdown of the density approximation as the threshold parameter approaches the smallest observation. The unboundedness can be eliminated with the correct likelihood $\mathcal{L}(\mu, \sigma, \gamma)$ as shown in Figure 1 (b).

The three-parameter gamma and Weibull distributions also fall into this category. Cheng and Traylor (1995) point out that one can extend the three-parameter Weibull distribution

to the maximum generalized extreme value (GEV) distribution by letting the power parameter become negative. Hirose (1996) gives details about how to obtain the minimum GEV distribution by reparameterizing the three-parameter Weibull distribution. The minimum GEV family has the cdf

$$G(x) = 1 - \exp \left\{ - \left[1 - \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\},$$

$$1 - \xi \left(\frac{x - \mu}{\sigma} \right) \geq 0, \xi \neq 0. \quad (5)$$

Here ξ is a shape parameter and μ and σ are, respectively, location and scale parameters. Coles and Dixon (1999) suggest that when the number of observations is small (say less than 30), no matter what the initial values are in the numerical optimization algorithm, ML estimation of the maximum GEV parameters can fail to converge properly. In the minimum GEV distribution, for any location parameter μ , one can always find a path along which the values of σ and ξ change and the density-approximation log-likelihood increases without bound. There is a similar result for the maximum GEV distribution. To illustrate this behavior Figure 2 (a) plots the density-approximation log-likelihood for a simulated sample of $n = 20$ observations from a minimum GEV distribution with $\mu = -2.2$, $\sigma = 0.5$, and $\xi = -0.2$ as a function of μ for three different combinations of fixed σ and ξ . Instead of using the profile log-likelihood plot with respect to the location parameter μ that blows up at any μ , an alternative density log-likelihood plot is used to present the unbounded behavior. This plot indicates that when the shape parameter $\xi < -1$, as the location parameter μ approaches $x_{(1)} - \sigma/\xi$ and $\sigma > 0$ and $\xi < -1$ are fixed, the density-approximation log-likelihood increases without bound. When the shape parameter $-1 < \xi < 0$, as the location parameter μ approaches $x_{(1)} - \sigma/\xi$ and $\sigma > 0$ and $-1 < \xi < 0$ are fixed, the density-approximation log-likelihood decreases without bound. The local maximum is close to the true μ at -2.2 . If the shape parameter $\xi > 0$, as the location parameter μ approaches $x_{(20)} - \sigma/\xi$, the density-approximation log-likelihood again increases without bound. For all of these cases, the unboundedness can be eliminated by using the correct likelihood as shown by the profile log-likelihood plot in Figure 2 (b).

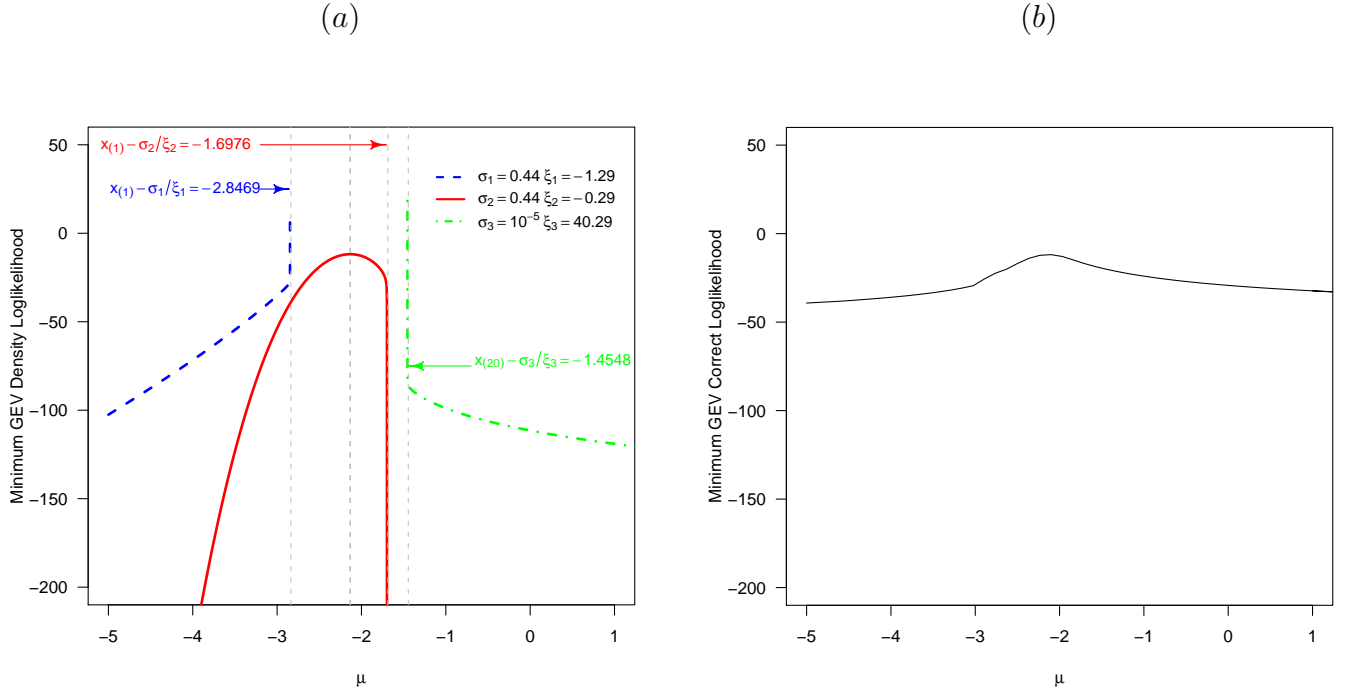


Figure 2: The minimum GEV log-likelihood plots of the location parameter μ for data with sample size $n = 20$ using (a) the density-approximation likelihood L and (b) the correct likelihood \mathcal{L} .

The Box-Cox (1964) transformation family of distributions with a location shift provides another example of an unbounded density-approximation likelihood in this category, as described in Chapters 6 and 9 of Atkinson (1985), Atkinson, Pericchi and Smith (1991), and Section 4 of Cheng and Traylor (1995). For a sample x_1, x_2, \dots, x_n , Box and Cox (1964) give the shifted power transformation as

$$y_i(x_i; \gamma, \lambda) = \begin{cases} \frac{(x_i + \gamma)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log(x_i + \gamma), & \text{if } \lambda = 0. \end{cases}$$

Suppose that $y_i \stackrel{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2), i = 1, \dots, n$. Then the likelihood function for the original observations $x_i, i = 1, \dots, n$ using the density approximation is (1) where the density is given by

$$f(x_i; \boldsymbol{\theta}) = \begin{cases} \frac{1}{\sigma} \phi_{\text{nor}} \left(\frac{(x_i + \gamma)^\lambda - \mu\lambda - 1}{\sigma\lambda} \right) |(x_i + \gamma)^{\lambda-1}|, & \text{if } \lambda \neq 0; \\ \frac{1}{\sigma|x_i + \gamma|} \phi_{\text{nor}} \left(\frac{\log(x_i + \gamma) - \mu}{\sigma} \right), & \text{if } \lambda = 0, \end{cases}$$

and $\boldsymbol{\theta} = (\mu, \sigma, \gamma, \lambda)'$.

Section 9.3 of Atkinson (1985) shows that the profile log-likelihood $\log L^*(\gamma) = \max_{\lambda, \mu, \sigma} \{\log L(\mu, \sigma, \gamma, \lambda)\}$ goes to ∞ as $\gamma \rightarrow -x_{(1)}$. Atkinson, Pericchi and Smith (1991) illustrate that the correct likelihood can be used to avoid the unbounded likelihood problem for this distribution.

2.2 Discrete Mixture Models Where at Least One Component Has Both a Location and a Scale Parameter

Suppose there are n iid observations x_1, x_2, \dots, x_n from the m -component discrete mixture distribution with the pdf

$$f(x; \boldsymbol{\theta}) = \sum_{i=1}^m p_i f_i(x; \boldsymbol{\theta}_i), \quad (6)$$

where $\boldsymbol{\theta} = (p_1, p_2, \dots, p_m, \boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \dots, \boldsymbol{\theta}'_m)'$, p_i is the proportion of component i with $\sum_{i=1}^m p_i = 1$ and for at least one i , the pdf for component i can be expressed as

$$f_i(x; \boldsymbol{\theta}_i) = \frac{1}{\sigma_i} \phi \left(\frac{x - \mu_i}{\sigma_i} \right).$$

That is, at least one component i belongs to the location-scale family with an unknown location parameter μ_i and scale parameter σ_i . Then if one sets a component location parameter equal to one of the observations, fixes the component proportion parameter at a positive value (less than 1), and allows the corresponding scale parameter to approach zero while fixing other parameter values, the likelihood increases without bound. Section 1.2.3 of Zucchini and MacDonald (2009) shows that replacing the density-approximation likelihood with the correct likelihood can again avoid the problem of unboundedness. Of course the same problem arises in mixtures of the corresponding log-location-scale distributions for which $\exp(\mu_i)$ is a scale parameter and σ_i is a shape parameter.

We use a simulated example to illustrate that replacing the density-approximation likelihood with the correct likelihood will eliminate the unbounded likelihood problem for the finite mixture models. We simulated one hundred observations x_1, x_2, \dots, x_{100} (following Example 1 of Yao 2010) from a mixture of two univariate normal components with proportions $p_1 = 0.7$, $p_2 = 0.3$, means $\mu_1 = 1$, $\mu_2 = 0$, and variances $\sigma_1^2 = 1$, $\sigma_2^2 = 0.25$. Let $\delta = \min\{\sigma_1, \sigma_2\} / \max\{\sigma_1, \sigma_2\}$, so that $\delta \in (0, 1]$. The original observations were rounded to one digit after the decimal point and thus the corresponding value of Δ used in the correct likelihood is 0.05. Figure 3 (a) shows that as δ approaches 0, the density-approximation profile log-likelihood of δ increases without bound. The counterpart in Figure 3 (b) using the correct likelihood solves the unboundedness problem. This plot also shows that the correct log-likelihood profile for δ tends to be flat for small values of δ . This is due to the fact that as δ approaches 0, the log-likelihood function is dominated by two parts: one part comes from the point mass at the smallest observation $x_{(1)}$, the other part is the log-likelihood of $x_{(2)}, \dots, x_{(100)}$ that follow the other normal distribution.

The switching regression model, described in Quandt (1972) and Quandt and Ramsey (1978), provides another example in this mixture category in which there is a finite mixture of regression models. For example, when there are $m = 2$ components, the pdf is given by (6) with

$$f_i(y; \boldsymbol{\theta}_i) = \frac{1}{\sigma_i} \phi_{\text{nor}} \left(\frac{y - \mathbf{x}_i' \boldsymbol{\beta}_i}{\sigma_i} \right).$$

The switching regression model is a special case of the m -component discrete mixture distribution with $\mu_i = \mathbf{x}_i' \boldsymbol{\beta}_i$.

Protheroe (1985) proposes a new statistic to describe the periodic ultra-high energy γ -ray signal source represented by a moving point on a circle. As explained there, events observed in time are caused by a background process in which the noise occurs according to a Poisson process with a constant intensity. Events from the “signal” occur according to a process with an intensity that is periodic with a known period P . For the event times T_1, T_2, \dots , the transformation $X_i = \text{mod}(T_i, P)/P$ maps the periodic event time into a circle with a unit circumference. As shown in Meeker and Escobar (1994), the pdf of X can be written as

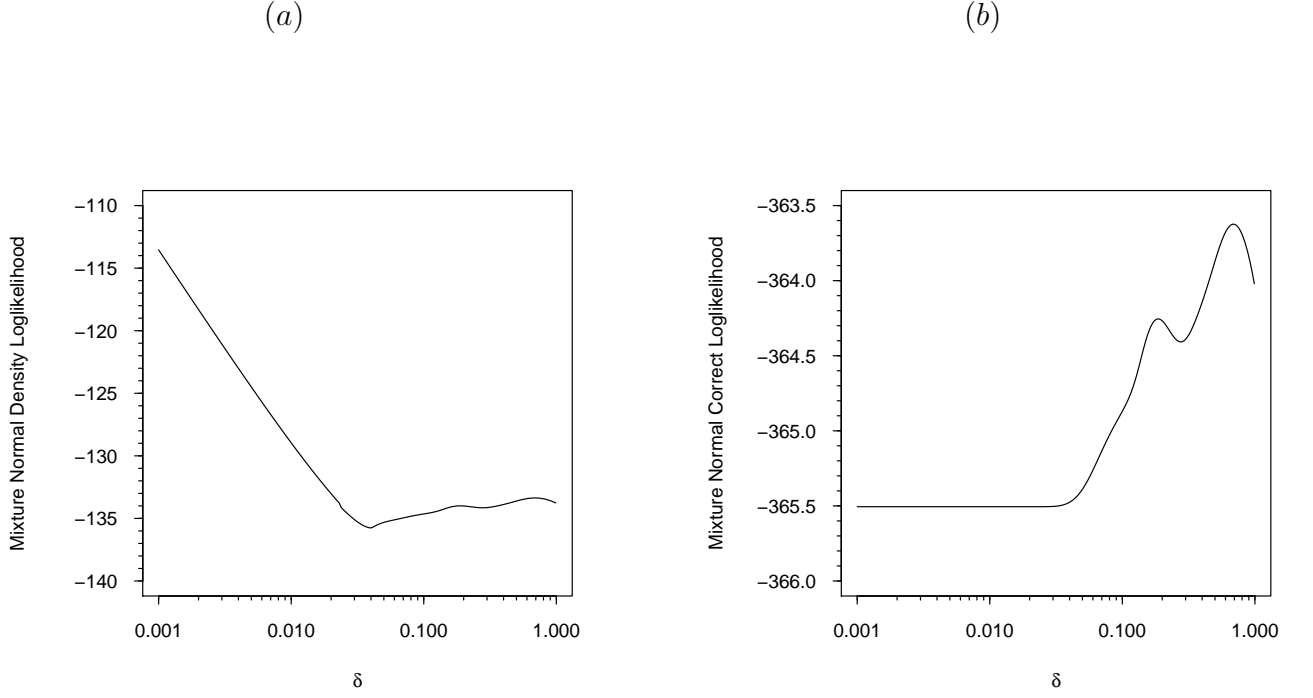


Figure 3: Mixture distribution of two univariate normal profile log-likelihood plots of δ using (a) the unbounded density-approximation likelihood and (b) the correct likelihood (with $\Delta=0.05$).

$$f_X(x; \boldsymbol{\theta}) = p + (1 - p)f_N(x; \mu, \sigma),$$

where $\boldsymbol{\theta} = (\mu, \sigma, p)'$, $0 \leq x \leq 1$, $0 < p < 1$, $0 \leq \mu \leq 1$, $\sigma > 0$, and

$$f_N(x; \mu, \sigma) = \frac{1}{\sigma} \sum_{j=-\infty}^{\infty} \phi_{\text{nor}} \left(\frac{x + j - \mu}{\sigma} \right),$$

where ϕ_{nor} is the standard normal pdf. Here, $f_N(x; \mu, \sigma)$ is the pdf for a “wrapped normal distribution.” For any $0 < p < 1$, let $\mu = x_i$ for any i , where x_1, x_2, \dots, x_n are iid transformed observations with the pdf $f_X(x; \boldsymbol{\theta})$, and then as $\sigma \rightarrow 0$, the product of the pdf’s approaches ∞ . Section 8.3.2 of Meeker and Escobar (1994) shows how to use ML to estimate the parameters of the wrapped normal distribution by using the bounded correct likelihood.

2.3 Minimum-Type (and Maximum-Type) Models Where at Least One of the Marginal Distributions Has Both a Location and a Scale Parameter

For $m \geq 2$ independent random variables X_1, X_2, \dots, X_m with cdf $F_i(x; \boldsymbol{\theta}_i)$ and pdf $f_i(x; \boldsymbol{\theta}_i)$, the cdf $F_{min}(x; \boldsymbol{\theta})$ of the minimum $X_{min} = \min\{X_1, X_2, \dots, X_m\}$ can be expressed as

$$F_{min}(x; \boldsymbol{\theta}) = \Pr(X_{min} \leq x) = 1 - \prod_{i=1}^m [1 - F_i(x; \boldsymbol{\theta}_i)]. \quad (7)$$

Then, the pdf of X_{min} is

$$f_{min}(x; \boldsymbol{\theta}) = \sum_{i=1}^m \left\{ f_i(x; \boldsymbol{\theta}_i) \prod_{j \neq i} [1 - F_j(x; \boldsymbol{\theta}_j)] \right\}.$$

Suppose again that for at least one i ,

$$f_i(x; \boldsymbol{\theta}_i) = \frac{1}{\sigma_i} \phi \left(\frac{x - \mu_i}{\sigma_i} \right),$$

which is the pdf belonging to the location-scale family with location parameter μ_i and scale parameter σ_i . Then for n iid observations x_1, x_2, \dots, x_n with the pdf $f_{min}(x; \boldsymbol{\theta})$, if one sets the location parameter for one of the components equal to the largest observation and allows the corresponding scale parameter to approach zero while fixing other parameter values, the likelihood increases without bound.

Friedman and Gertsbakh (1980) describe a minimum-type distribution (MTD) that is a special case of (7) with two random failure times: $t_A \sim \text{Weibull}(\alpha_A, \beta_A)$, $t_B \sim \text{Exp}(\alpha_B)$. The failure time of the device is $T = \min\{t_A, t_B\}$ with cdf

$$\Pr(T \leq t) = 1 - \exp \left[-\frac{t}{\alpha_B} - \left(\frac{t}{\alpha_A} \right)^{\beta_A} \right]. \quad (8)$$

The cdf for $Y = \log(T)$ can be written as

$$\Pr(Y \leq y) = 1 - \exp \left[-\exp(y - \mu_B) - \exp \left(\frac{y - \mu_A}{\sigma_A} \right) \right],$$

where $\mu_B = \log(\alpha_B)$ and $\mu_A = \log(\alpha_A)$ are location parameters and $\sigma_A = 1/\beta_A$ is the scale parameter of the smallest extreme value distribution. Friedman and Gertsbakh (1980) show that if all three parameters are unknown, there exists a path in the parameter space along which the likelihood function tends to infinity for any given sample and suggest an alternative method of estimation.

We simulated 100 observations t_1, t_2, \dots, t_{100} from the MTD in (8) with $\alpha_A = 2$, $\beta_A = 4$, and $\alpha_B = 1$. Figure 4 (a) shows that when $\mu_A = \log(t_{(100)})$, μ_B is fixed and σ_A approaches zero, the log-likelihood function increases without bound. The unboundedness problem can again be resolved by using the correct likelihood as shown in Figure 4 (b).

The correct likelihood has a global maximum in situations where a sufficient amount of data is available from each component of the minimum process. When, however, one or the other of the minimum process dominates in generating the data, due to particular values of the parameters or right censoring, there can be an identifiability problem so that a unique maximum of the three-parameter likelihood will not exist.

The simple disequilibrium model illustrated in Griliches and Intriligator (1983) is another special case of the minimum-type model. They consider two random variables X_1 and X_2 from a normal distribution leading again to the likelihood function in (1) with

$$\begin{aligned} f(y_i; \boldsymbol{\theta}) &= \frac{1}{\sigma_1} \phi_{\text{nor}} \left(\frac{y_i - \mathbf{x}'_{1i} \boldsymbol{\beta}_1}{\sigma_1} \right) \left[1 - \Phi_{\text{nor}} \left(\frac{y_i - \mathbf{x}'_{2i} \boldsymbol{\beta}_2}{\sigma_2} \right) \right] \\ &+ \frac{1}{\sigma_2} \phi_{\text{nor}} \left(\frac{y_i - \mathbf{x}'_{2i} \boldsymbol{\beta}_2}{\sigma_2} \right) \left[1 - \Phi_{\text{nor}} \left(\frac{y_i - \mathbf{x}'_{1i} \boldsymbol{\beta}_1}{\sigma_1} \right) \right], \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \sigma_1, \sigma_2)'$. An argument very similar to that employed in the case of the switching regression model can be used to show that the simple disequilibrium model has an unbounded likelihood function. Again, using the correct likelihood avoids this problem.

Maximum-type models for which at least one of the marginal distributions has both a location and scale parameter have the same problem as the minimum-type models.

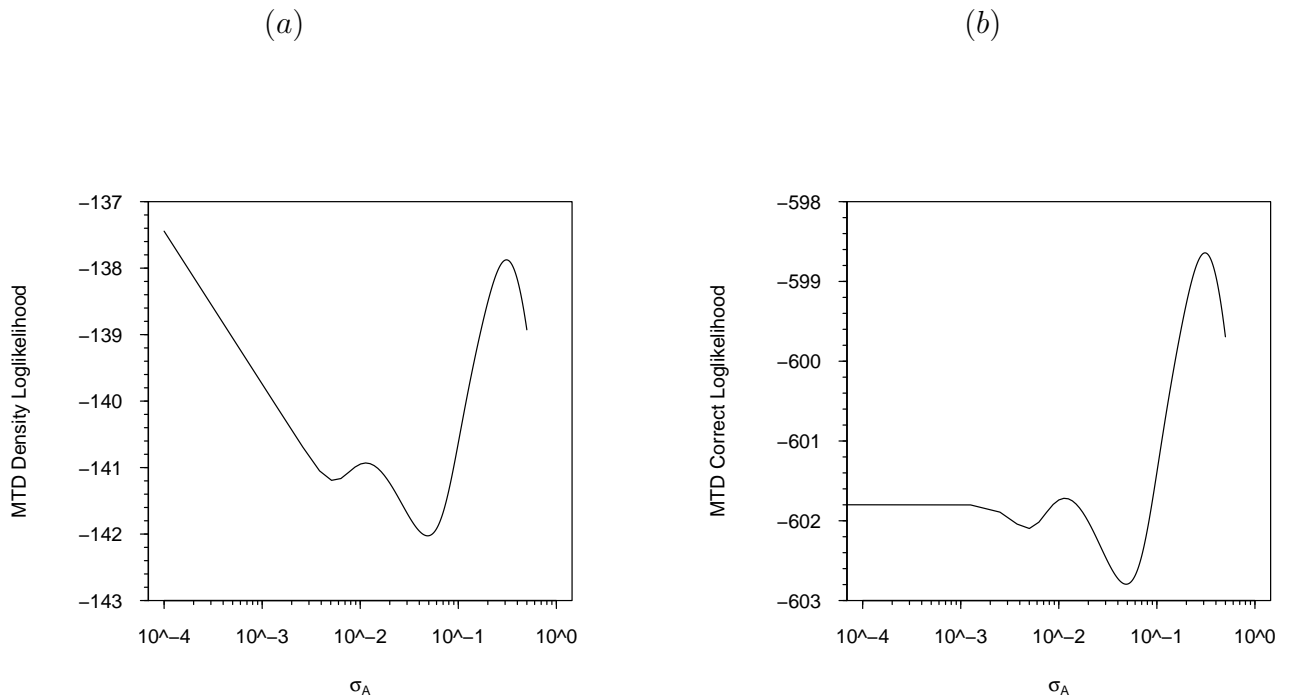


Figure 4: An MTD corresponding to a minimum of two independent random variables having the Weibull and exponential distributions. The profile log-likelihood plots of σ_A using (a) the density-approximation likelihood and (b) the correct likelihood (with $\Delta=0.05$).

3 The Effect of the Round-off Error on Estimation

In this section we use two examples to explore the effect that the round-off error Δ has on estimation with the correct likelihood. Of course, if one knows the precision of the measuring instrument and the rule that was used for rounding one's data, the choice of Δ is obvious. Sometimes, however, the precision of the measuring instrument and the exact rounding scheme that was used for a data set are unknown.

3.1 Background

Giesbrecht and Kempthorne (1976) present an empirical study that investigates the effect of the round-off error Δ for estimation of the parameters for the three-parameter lognormal distribution. As Δ decreases, the asymptotic variances and covariances of the MLE of the

parameters rapidly approach the values for the no-censoring case provided in Harter and Moore (1965). Atkinson, Pericchi and Smith (1991) suggest that care is needed in choosing Δ and recommend that one could examine profile likelihood plots for different values of Δ and choose the value of Δ that makes the profile likelihood smooth near $\mu = -y_{(1)}$ for the Box-Cox (1964) shifted power transformation model. Vardeman and Lee (2005) discuss how to use the likelihood function based on the rounded observations to make inferences about the parameters of the underlying distribution. They emphasize that the relationship between the range of the rounded observations and the rounding rule, defined by Δ , can have an important effect on inferences when the ratio of Δ/σ is large (say more than 3).

3.2 Numerical Examples

In the first example, we investigate the effect of the round-off error on estimation of the location parameter for the minimum GEV distribution. Vardeman and Lee (2005) suggest (for a different distribution) that the round-off error Δ should be sufficiently large (equivalently, the number of digits after the decimal point should be sufficiently small), relative to the range of the rounded observations, so that the maximum of the correct likelihood function exists. We consider four different minimum GEV distributions with a common location parameter $\mu = -2.2$ and shape parameter $\xi = -0.2$ but different scale parameters $\sigma = 0.3, 0.5, 1$, and 2 . For each minimum GEV distribution, we generated 15 samples, each with $n = 20$ observations. For a given rounding scheme, the correct likelihood of the minimum GEV distribution (proportional to the probability of the data) for a sample x_1, x_2, \dots, x_{20} can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &\propto \prod_{i=1}^{20} \Pr(l_{x_{i_k}} < x_i < u_{x_{i_k}}) \\ &= \prod_{i=1}^{20} \left[G(u_{x_{i_k}}) - G(l_{x_{i_k}}) \right] \\ &= \prod_{i=1}^{20} \left(\exp \left\{ - \left[1 - \xi \left(\frac{l_{x_{i_k}} - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} - \exp \left\{ - \left[1 - \xi \left(\frac{u_{x_{i_k}} - \mu}{\sigma} \right) \right]^{-1/\xi} \right\} \right), \end{aligned}$$

where $l_{x_{i_k}}$ and $u_{x_{i_k}}$ are respectively the lower and upper endpoints of the interval with k

digits after the decimal point for the i th observation and $\boldsymbol{\theta} = (\mu, \sigma, \xi, k)'$.

For different values of σ , the plots in Figure 5 compare the location parameter estimates $\hat{\mu}$ when k , the number of digits after the decimal point, varies from 1 to 4 (i.e., the corresponding round-off error Δ changes from 0.05 to 0.00005). We did not use the same y-axis range in these four plots to allow focus on the stability of the estimates as a function of k . Figure 5 indicates that when σ is small, more samples provide different parameter estimates $\hat{\mu}$ for different numbers of digits after the decimal point k than in the situation with large σ . This is true especially when the number of digits after the decimal point k is small (i.e., Δ is large). In this example, when σ is 2, the choice of k does not significantly affect the estimation of the location parameter μ .

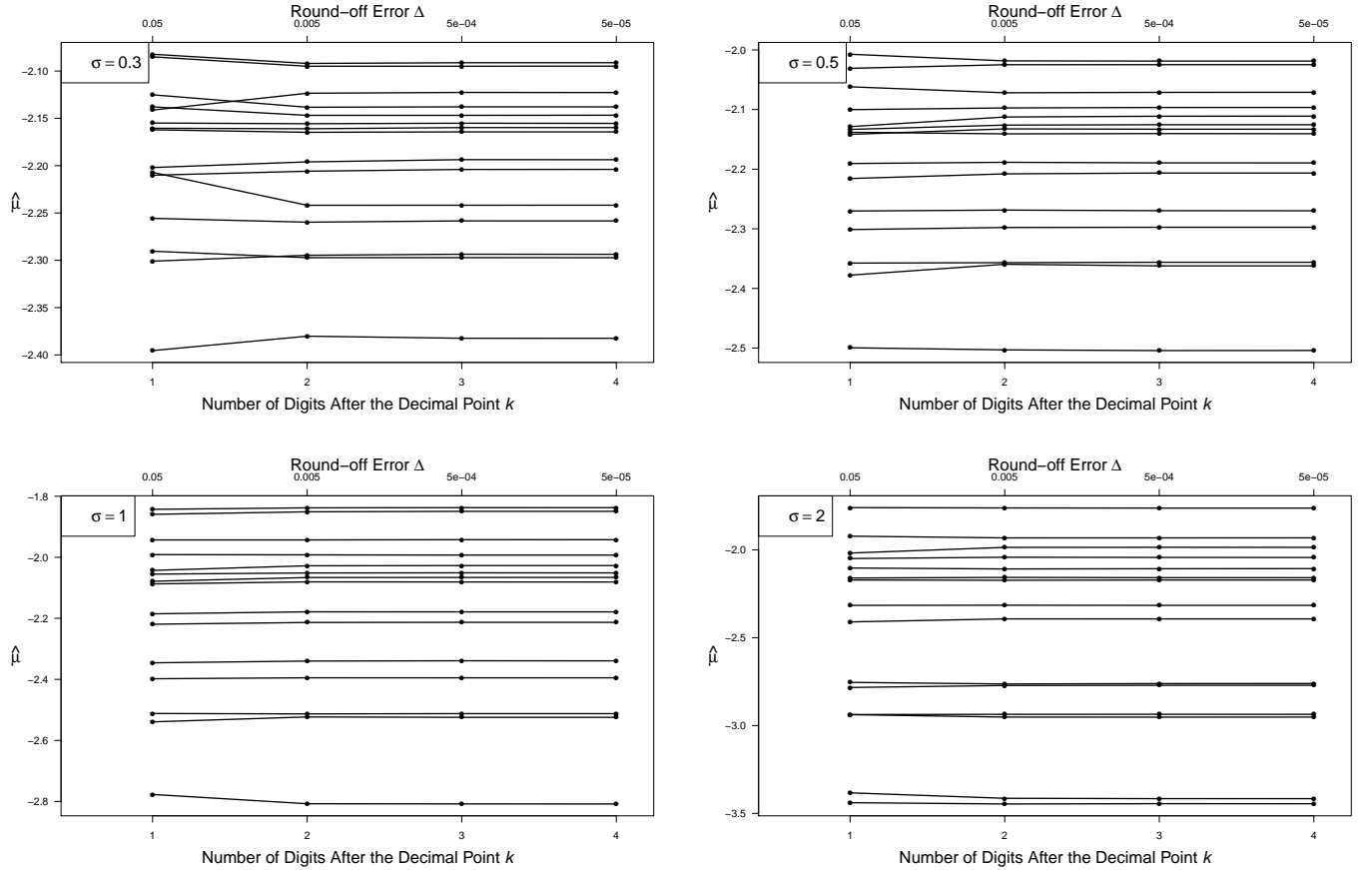


Figure 5: Plots of the minimum GEV location parameter estimates $\hat{\mu}$ versus the number of digits after the decimal point k with σ equal to 0.3, 0.5, 1, and 2. The Δ values on the top of each plot correspond to the round-off errors.

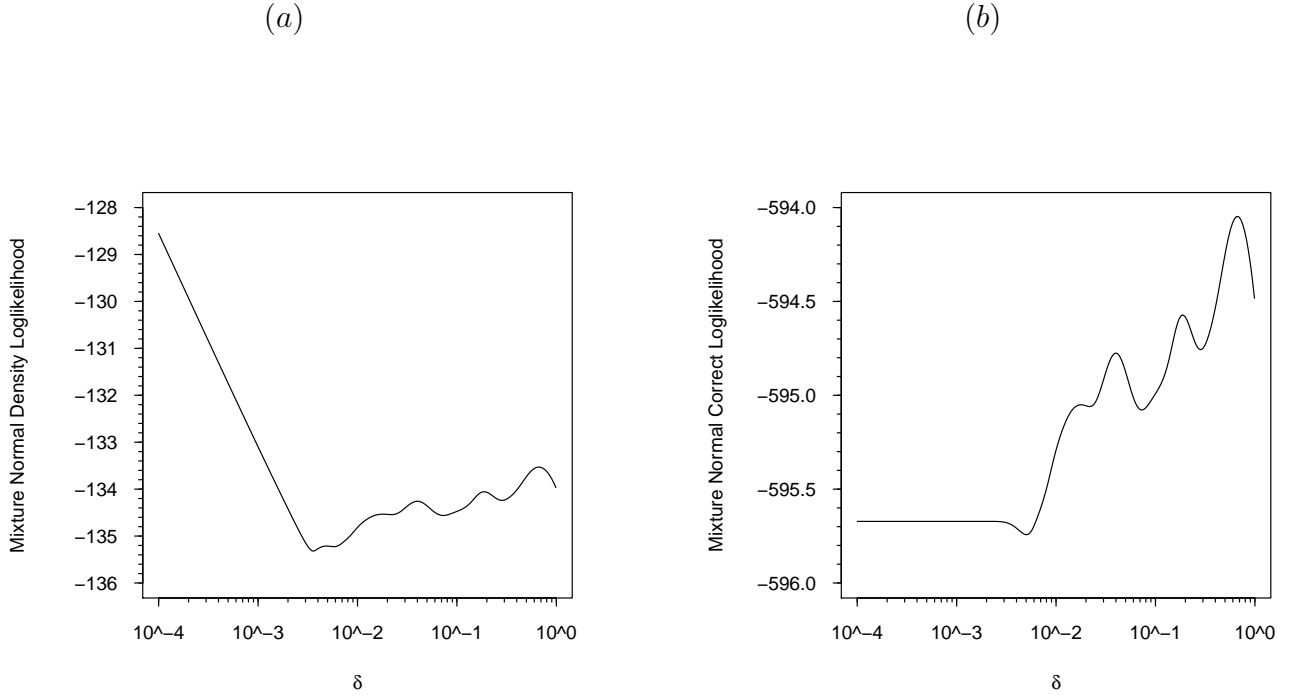


Figure 6: Mixture distribution of two univariate normal profile log-likelihood plots of δ using (a) the unbounded density-approximation likelihood and (b) the correct likelihood (with $\Delta = 0.005$).

For the second example, we return to the mixture of two univariate normal distributions in Section 2.2 to study the effect that Δ in the correct likelihood has on estimation. The plots in Figure 3 for $\Delta = 0.05$ are relatively smooth. Rounding the same original observations x_1, x_2, \dots, x_{100} to two digits after the decimal point, the profile log-likelihood functions with $\Delta = 0.005$ in Figure 6 are more wiggly than the corresponding profile log-likelihood functions, in Figure 3 with $\Delta = 0.05$, especially in the right part of the plots. The occurrence of multiple bumps in the profile log-likelihood curves, for $\Delta = 0.005$ is due to the fact that increasing the data precision will result in more distinct clusters with data points close together. As in Figure 3, we note that the profile log-likelihood plot using the correct likelihood is flat as δ approaches 0.

4 A Sufficient Condition for Using the Density-Approximation Likelihood

As discussed in Sections 1.2 and 1.3, the likelihood function based on the usual density approximation for n iid observations t_i , $i = 1, \dots, n$ has the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n L_i(\boldsymbol{\theta}; t_i) = \prod_{i=1}^n f(t_i; \boldsymbol{\theta}).$$

The correct likelihood based on small intervals $(t_i - \Delta, t_i + \Delta)$ (implied by the data's precision) can be expressed as

$$\mathcal{L}_\Delta(\boldsymbol{\theta}) = \prod_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}; t_i) = \left(\frac{1}{2\Delta}\right)^n \prod_{i=1}^n \int_{t_i-\Delta}^{t_i+\Delta} f(t; \boldsymbol{\theta}) dt.$$

Here, for simplicity we assume that $\Delta_i = \Delta$ for all i . Kempthorne and Folks (1971, page 259) state that a sufficient condition for $L(\boldsymbol{\theta})$ to be proportional to the probability of the data (i.e., to be a proper likelihood) is to satisfy the Lipschitz condition which states that for all $\epsilon \neq 0$ in the interval $(-\Delta/2, \Delta/2)$ ($\Delta > 0$ is fixed), there exists a function $h(t; \boldsymbol{\theta})$ such that

$$\left| \frac{f(t + \epsilon; \boldsymbol{\theta}) - f(t; \boldsymbol{\theta})}{\epsilon} \right| < h(t; \boldsymbol{\theta}).$$

Kempthorne and Folks (1971) did not provide a proof and we have been unable to find one. Also, we have been unable to construct or find a pdf $f(t; \boldsymbol{\theta})$ that satisfies the Lipschitz condition but has different ML estimates when using the density approximation and the correct likelihoods as $\Delta \rightarrow 0$. We, therefore, look at this problem from a different perspective and provide an alternative sufficient condition through Theorem 1 and its corollary.

Theorem 1: For n iid observations t_1, t_2, \dots, t_n from a distribution with the pdf $f(t; \boldsymbol{\theta})$ and $\boldsymbol{\theta} \in \Theta \subset R^p$, if $\{f(t; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$ is equicontinuous at t_1, t_2, \dots, t_n , then the correct likelihood $\{\mathcal{L}_\Delta(\boldsymbol{\theta})\}$ converges uniformly to the density-approximation likelihood $L(\boldsymbol{\theta})$ as $\Delta \rightarrow 0$.

The proof is given in the Appendix 6.1. By the property of a uniformly convergent sequence described by Intriligator (1971), if $\{\mathcal{L}_\Delta(\boldsymbol{\theta})\}$ converges uniformly to $L(\boldsymbol{\theta})$ as $\Delta \rightarrow 0$,

then $\{\boldsymbol{\theta}_\Delta^*\}$ converges to $\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}_\Delta^*$ and $\boldsymbol{\theta}^*$ are the unique maximizers of $\mathcal{L}_\Delta(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$, respectively, assuming that $\boldsymbol{\theta}_\Delta^*$ and $\boldsymbol{\theta}^*$ exist in the parameter space Θ . We now have the following corollary.

Corollary 1: For n iid observations t_1, t_2, \dots, t_n from a distribution with the pdf $f(t; \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta \subset R^p$, if $\{f(t; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$ is equicontinuous at t_1, t_2, \dots, t_n , then $\{\boldsymbol{\theta}_\Delta^*\}$ converges to $\boldsymbol{\theta}^*$ as $\Delta \rightarrow 0$, where $\boldsymbol{\theta}_\Delta^*$ and $\boldsymbol{\theta}^*$ are the unique maximizers of $\mathcal{L}_\Delta(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$, respectively, assuming that $\boldsymbol{\theta}_\Delta^*$ and $\boldsymbol{\theta}^*$ exist in the parameter space Θ .

The equicontinuity condition, however, is not necessary. An example is given in the Appendix 6.2.

5 Concluding Remarks

In this paper, we used the correct likelihood based on small intervals instead of the density approximation to eliminate the problem of an unbounded likelihood. We explored several classes of models where the unbounded “likelihood” arises when using the density-approximation likelihood and illustrated how using a correctly expressed likelihood can eliminate the problem. We investigated the effect that the round-off error has on estimation with the correct likelihood, especially under the circumstance having unknown precision of the measuring instrument. The equicontinuity condition is sufficient for the density approximation and correct likelihoods to provide the same MLE’s as the round-off error $\Delta \rightarrow 0$.

6 Appendix

6.1 Proof of Theorem 1

If $\{f(t; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$ is equicontinuous at t_1, t_2, \dots, t_n , then for each t_i , for any $\epsilon > 0$, there exists a $\delta > 0$ such that if $|t - t_i| < \delta$, then $|f(t; \boldsymbol{\theta}) - f(t_i; \boldsymbol{\theta})| < \epsilon$ for all $\boldsymbol{\theta} \in \Theta$. Hence, for $0 < \Delta < \delta$,

$$\begin{aligned}
\left| \frac{1}{2\Delta} \int_{t_i-\Delta}^{t_i+\Delta} f(t; \boldsymbol{\theta}) dt - f(t_i, \boldsymbol{\theta}) \right| &= \left| \frac{1}{2\Delta} \int_{t_i-\Delta}^{t_i+\Delta} [f(t; \boldsymbol{\theta}) - f(t_i; \boldsymbol{\theta})] dt \right| \\
&\leq \frac{1}{2\Delta} \int_{t_i-\Delta}^{t_i+\Delta} |f(t; \boldsymbol{\theta}) - f(t_i; \boldsymbol{\theta})| dt \\
&< \epsilon.
\end{aligned}$$

Therefore, for every t_i , $\{\frac{1}{2\Delta} \int_{t_i-\Delta}^{t_i+\Delta} f(t; \boldsymbol{\theta}) dt\}$ converges uniformly to $f(t_i, \boldsymbol{\theta})$, and thus $\{\mathcal{L}_\Delta(\boldsymbol{\theta})\}$ converges uniformly to $L(\boldsymbol{\theta})$ as $\Delta \rightarrow 0$.

6.2 An Example Showing the Equicontinuity Condition Is Not Necessary

The pdf $\{f(x; \theta)\}_{\theta \in \Theta}$ below is not equicontinuous at $x = 0$ or 2 , but the ML estimates using the density-approximation likelihood and the correct likelihood are the same as $\Delta \rightarrow 0$. For $0 < \theta < 2$, the univariate pdf has the form

$$f(x, \theta) = \begin{cases} 1 - \frac{x}{\theta}, & \text{if } 0 < x < \theta; \\ \frac{x-\theta}{2-\theta}, & \text{if } \theta \leq x < 2. \end{cases}$$

Also, $f(x, 0) = x/2$ for $0 \leq x \leq 2$ and $f(x, 2) = 1 - x/2$ for $0 \leq x \leq 2$. Note that $\{f(x, \theta)\}_{\theta \in \Theta}$ is not equicontinuous at $x = 0$ or 2 , because $f(x, \theta)$ is not even continuous at $x = 0$ or 2 for $\theta \in [0, 2]$. Suppose that $n = 1$ and we have one observation x_1 . When using the density-approximation likelihood, if $0 \leq x_1 < 1$, the ML estimate is $\hat{\theta} = 2$; if $1 < x_1 \leq 2$, the ML estimate is $\hat{\theta} = 0$; if $x_1 = 1$, the ML estimate is $\hat{\theta} = 0$ or 2 (not unique). For $0 < \theta < 2$, the cdf has the form

$$F(x, \theta) = \begin{cases} (2 - \frac{x}{\theta}) \frac{x}{2}, & \text{if } 0 \leq x < \theta; \\ \frac{\theta}{2} + \frac{(x-\theta)^2}{2(2-\theta)}, & \text{if } \theta \leq x \leq 2. \end{cases}$$

Also, $F(x, 0) = x^2/4$ for $0 \leq x \leq 2$, and $F(x, 2) = x - x^2/4$ for $0 \leq x \leq 2$.

For $0 < x_1 < 2$, consider $0 < \Delta < \min\left(\frac{x_1}{8}, \frac{2-x_1}{8}\right)$. Then the correct likelihood is

$$\begin{aligned}\mathcal{L}_\Delta(\theta) &= \frac{1}{2\Delta} [F(x_1 + \Delta, \theta) - F(x_1 - \Delta, \theta)] \\ &= \begin{cases} \frac{x_1}{2}, & \text{if } \theta = 0; \\ \frac{x_1 - \theta}{2 - \theta}, & \text{if } 0 < \theta < x_1 - \Delta; \\ \frac{1}{4\Delta} \left[\frac{(x_1 + \Delta - \theta)^2}{2 - \theta} + \frac{(x_1 - \Delta - \theta)^2}{\theta} \right], & \text{if } x_1 - \Delta \leq \theta \leq x_1 + \Delta; \\ 1 - \frac{x_1}{\theta}, & \text{if } x_1 + \Delta < \theta < 2; \\ 1 - \frac{x_1}{2}, & \text{if } \theta = 2. \end{cases}\end{aligned}$$

Note that for $x_1 - \Delta \leq \theta \leq x_1 + \Delta$,

$$\begin{aligned}\mathcal{L}_\Delta(\theta) &\leq \frac{1}{4\Delta} \left[\frac{(2\Delta)^2}{2 - \theta} + \frac{(2\Delta)^2}{\theta} \right] \leq \frac{\Delta}{(2 - x_1) - \Delta} + \frac{\Delta}{x_1 - \Delta} \\ &\leq \frac{\Delta}{8\Delta - \Delta} + \frac{\Delta}{8\Delta - \Delta} = \frac{2}{7};\end{aligned}$$

for $0 < \theta < x_1 - \Delta$,

$$\mathcal{L}_\Delta(\theta) < \frac{x_1 - \frac{x_1}{2}\theta}{2 - \theta} = \frac{x_1}{2} = \mathcal{L}_\Delta(0);$$

and for $x_1 + \Delta < \theta < 2$,

$$\mathcal{L}_\Delta(\theta) < 1 - \frac{x_1}{2} = \mathcal{L}_\Delta(2).$$

Thus the ML estimate using the correct likelihood is

$$\hat{\theta}_\Delta = \begin{cases} 2, & \text{if } 0 < x_1 < 1; \\ 0, & \text{if } 1 < x_1 < 2; \\ 0 \text{ or } 2, & \text{if } x_1 = 1. \end{cases}$$

This is exactly the same as the $\hat{\theta}$ we obtained before when using the density-approximation likelihood.

For $x_1 = 0$, we have

$$\mathcal{L}_\Delta(\theta) = \frac{1}{2\Delta} F(\Delta, \theta)$$

$$= \begin{cases} \frac{\Delta}{8}, & \text{if } \theta = 0; \\ \frac{1}{2\Delta} \left[\frac{\theta}{2} + \frac{(\Delta-\theta)^2}{2(2-\theta)} \right], & \text{if } 0 < \theta \leq \Delta; \\ \frac{1}{2} - \frac{\Delta}{4\theta}, & \text{if } \Delta < \theta < 2; \\ \frac{1}{2} - \frac{\Delta}{8}, & \text{if } \theta = 2. \end{cases}$$

Thus, $\hat{\theta}_\Delta = 2$ (for $0 < \Delta < \frac{1}{8}$), the same as $\hat{\theta}$. Similarly, for $x_1 = 2$, we also have $\hat{\theta}_\Delta = \hat{\theta} = 0$ (for $0 < \Delta < \frac{1}{8}$).

References

- [1] Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Oxford: Oxford University Press.
- [2] Atkinson, A. C., Pericchi, L. R. and Smith, R. L. (1991). Grouped likelihood for the shifted power transformation. *Journal of the Royal Statistical Society: Series B*, 53: 473-482.
- [3] Barnard, G. A. (1967). The use of the likelihood function in statistical practice. *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 27-40.
- [4] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26: 211-252.
- [5] Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society: Series B*, 45: 394-403.
- [6] Cheng, R. C. H. and Iles, T. C. (1987). Corrected maximum likelihood in non-regular problems. *Journal of the Royal Statistical Society: Series B*, 49: 95-101.

- [7] Cheng, R. C. H. and Traylor, L. (1995). Non-regular maximum likelihood problems. *Journal of the Royal Statistical Society: Series B*, 57: 3-44.
- [8] Coles, S. G. and Dixon, M.J. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2: 5-23.
- [9] Fisher, R. A (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41: 155-160.
- [10] Fisher, R. A (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London: Series A*, 222: 309-368.
- [11] Friedman, L. and Gertsbakh, I. B. (1980). Maximum likelihood estimation in a minimum-type model with exponential and Weibull failure modes. *Journal of the American Statistical Association*, 75: 460-465.
- [12] Giesbrecht, F. and Kempthorne, O. (1976). Maximum likelihood estimation in the three-parameter lognormal distribution. *Journal of the Royal Statistical Society: Series B*, 38: 257-264.
- [13] Griliches, Z. and Intriligator, M. D. (1983). *Handbook of Econometrics*. Amsterdam, New York: North-Holland Publishing.
- [14] Harter, H. L. and Moore, A. H. (1965). Maximum-likelihood estimation of the parameters of gamma and Weibull populations from complete and from censored samples. *Technometrics*, 7: 639-643.
- [15] Hirose, H. (1996). Maximum likelihood estimation in the 3-parameter Weibull distribution: A look through the generalized extreme-value distribution. *IEEE Transactions on Dielectrics and Electrical Insulation*, 3: 43-55.
- [16] Intriligator, M. D. (1971). *Mathematical Optimization and Economic Theory*. Englewood Cliffs, N.J.: Prentice-Hall.
- [17] Kempthorne, O. (1966). Some aspects of experimental inference. *Journal of the American Statistical Association*, 61: 11-34.

- [18] Kempthorne, O. and Folks, L. (1971). *Probability, Statistics, and Data Analysis*. Ames, Iowa: Iowa State University Press.
- [19] Kulldorff, G. (1957). On the conditions for consistency and asymptotic efficiency of maximum likelihood estimate, *Skandinavisk Aktuarie Tidskrift*, 40: 129-144.
- [20] Kulldorff, G. (1961). *Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples*, Stockholm, Sweden: Almqvist and Wiksell.
- [21] Meeker, W. Q. and Escobar, L. A. (1994). Maximum likelihood methods for fitting parametric statistical models. In Stanford, J. L. and Vardeman, S. B. (Eds.), *Methods in Experimental Physics*, pp. 211-244. New York: Academic Press.
- [22] Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley and Sons.
- [23] Nelson, W. B. (1982). *Applied Life Data Analysis*. New York: John Wiley and Sons.
- [24] Protheroe, R. J. (1985). A new statistic for the analysis of circular data with applications in ultra-high energy gamma-ray astronomy. *Astronomy Express*, 1: 137-142.
- [25] Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67: 306-310.
- [26] Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73: 730-738.
- [27] Smith, R. L. and Weissman, I. (1985). Maximum likelihood estimation of the lower tail of a probability distribution. *Journal of the Royal Statistical Society B*, 47: 285-298.
- [28] Vardeman, S. B. (2005). Lecture notes: examples of a mixture of two uniform distributions. <http://www.public.iastate.edu/~vardeman/stat543/Lectures/Stat543-2-4-05.pdf>, downloaded January 7, 2013.
- [29] Wong, T. S. T. and Li, W. K. (2006). A note on the estimation of extreme value distributions using maximum product of spacings. In *Time Series and Related Topics*:

In Memory of Ching-Zong Wei, pp. 272-283. Beachwood, Ohio: Institute of Mathematical Statistics.

- [30] Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*, 140: 2089-2098.
- [31] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton: Chapman and Hall/CRC.