Stat Papers (2009) 50:195–202
DOI 10.1007/s00362-007-0056-5

NOTE

On a criticism of the profile likelihood function

José A. Montoya · Eloísa Díaz-Francés · David A. Sprott

Received: 6 March 2006 / Revised: 22 January 2007 / Published online: 1 March 2007

Abstract The profile likelihood function is often criticized for giving strange or unintuitive results. In the cases discussed here these are due to the use of density functions that have singularities. These singularities are naturally inherited by the profile likelihood function. It is therefore apparently important to be reminded that likelihood functions are proportional to probability functions, and so cannot have singularities. When this issue is addressed, then the profile likelihood poses no problems of this sort. This is of particular importance since the profile likelihood is a commonly used method for dealing with separate estimation of parameters.

Keywords Conditional likelihood · Integrated likelihood · Marginal likelihoods · Maximized likelihood · Nuisance parameters · Singularities · Threshold parameters

1 Introduction

The profile or maximized likelihood function is a powerful though simple method devised to handle estimation of parameters of interest in the presence of nuisance parameters. For this purpose, it is more generally available than are alternative likelihoods such as the conditional, marginal, or integrated likelihoods, which depend on special structures and hence are more restrictive.

Let Y be a discrete random variable with probability function $f(y;\theta,\lambda)$ in terms of the unknown parameters θ, λ. The parameter of interest is θ and λ

J. A. Montoya · E. Díaz-Francés (✉) · D. A. Sprott
Centro de Investigación en Matemáticas, A.P. 402, Guanajuato, Gto. 36000 Mexico, Mexico
e-mail: diazfran@cimat.mx

is a so-called nuisance parameter. The profile likelihood and its corresponding relative likelihood function of θ, standardized to be one at the maximum of the likelihood function, are defined for a sample of independent observations $y = (y_1, \ldots, y_n)$ as

$$L_{\max}(\theta; y) \propto \prod_{i=1}^{n} f\left[y_i; \theta, \hat{\lambda}(\theta, y)\right],$$
$$R_{\max}(\theta; y) \propto \prod_{i=1}^{n} f\left[y_i; \theta, \hat{\lambda}(\theta, y)\right] \Big/ \sup\left\{\prod_{i=1}^{n} f(y_i; \theta, \lambda)\right\}, \qquad (1)$$

so that $0 \leq R_{\max} \leq 1$, where $\hat{\lambda}(\theta, y)$ is the restricted maximum likelihood estimate (mle) of λ for a specified value of θ. If $\hat{\theta}$ is the mle of θ, it follows that $\hat{\lambda}(\hat{\theta}) \equiv \hat{\lambda}$, the unrestricted or overall mle of λ. Since f is a probability function, it is necessarily bounded and so the denominator in (1) exists and is finite.

If Y is a continuous random variable, f is usually taken to be the corresponding density function in (1). There can be a difficulty with this because a density function can have a singularity and hence be unbounded. Then the denominator in (1) may be infinity and so $R_{\max}(\theta; y)$ of (1) is undefined. This can lead to various “strange”profile likelihood functions that make no scientific sense, Example 3.2, and (1) is often criticized for this behavior. Since the profile likelihood is perhaps the most readily available method for dealing with nuisance parameters, it is important to examine whether such criticisms are justifiable, or whether they are merely the result of a mathematical error.

However data y must in reality *always* be discrete since all measuring instruments have finite precision, and the data can only be recorded to a finite number of decimals. Thus the observation $Y = y$ can be interpreted as $y - \frac{1}{2}h \leq Y \leq y + \frac{1}{2}h$, where h is the precision of the measuring instrument, and so is a fixed positive number. Therefore, for independent observations $y = (y_1, \ldots, y_n)$, the resulting likelihood function is proportional to the probability function

$$L_{\max}(\theta; y) \propto \prod_{i=1}^{n} P(y_i - \frac{1}{2}h \leq Y_i \leq y_i + \frac{1}{2}h)$$
$$= \prod_{i=1}^{n} \int_{Y_i = y_i - \frac{1}{2}h}^{y_i + \frac{1}{2}h} f\left[Y_i; \theta, \hat{\lambda}(\theta, y)\right] dY_i. \qquad (2)$$

Allowing $h = 0$ implies the measuring instrument has infinite precision and that the observations can be recorded to an infinite number of decimals. Since for a continuous random variable $Y, P(Y = y; \theta) = 0$ for all y and θ, this cannot be the basis for obtaining a likelihood function. If in contrast, one assumes that the precision of the measuring instrument is $h > 0$, then conditions are required for the density function $f(y;\theta)$ to be used as an approximation to the likelihood function (2). But if the density function has a singularity at any given value of θ, then these conditions are violated and $f(y;\theta)$ cannot be used to approximate the likelihood function at that value of θ.

2 Continuous approximations

The representation of observations by continuous random variables involves an approximation. Proper attention to this approximation is necessary if the associated probability density function is to replace the probability function $f(y;\theta)$ in (1), Barnard and Sprott (1983). It is unfortunate that this approximating process is usually ignored and the density function taken for granted.

The approximating process leading to the use of the density function can be summarized as follows. By the mean value theorem for integrals of continuous functions, the ith integral in (2) is $hf(y';\theta)$ for some $y' \in \left[y_i - \frac{1}{2}h, y_i + \frac{1}{2}h\right]$. If $f(y;\theta)$ is approximately constant in this range for all plausible θ, then $f(y';\theta) \approx f(y;\theta)$ in this same range. If this approximation is adequate for all $i \in \{1,\ldots,n\}$ and if h does not depend on θ, the density function $f(y;\theta)$ can replace the probability function in (1).

The necessity of such an argument has been discussed, for example, by Kalbfleisch (1985, Sect. 9.4), Edwards (1992, p. 6, p. 167), Lindsey (1998), Sprott (2000, p. 19, pp. 203–294), Lawless (2003, p. 186), Meeker and Escobar (1998, p. 275). It is important to emphasize that (2) is not an ad hoc attempt to discretize a continuous random variate y with density function $f(y)$ which is considered the fundamental entity. It is quite the opposite, a continuous approximation to a discrete random variate y with probability function $f(y)$ which is the fundamental entity. The main purpose of the continuous approximation is for mathematical convenience; derivatives and integrals are easier to handle than finite differences and sums.

3 Examples

Example 3.1 A well known example is that of a single observation y from the $N(\lambda,\sigma^2)$ distribution.

Using the continuous approximation, the "profile" likelihood function of σ produced by $Y = y$, $\hat{\lambda}(\sigma) = y$ is proportional to σ^{-1}, which is not defined at the mle $\hat{\sigma} = 0$. Thus the relative profile likelihood is undefined.

In contrast, if one takes into account that the observation is in reality discrete and that the precision of the measuring instrument is h, a fixed positive number, using (2) with $\hat{\lambda} = y$, the resulting profile likelihood function is

$$
\begin{aligned}
L_{\max}(\sigma; y) &\propto \int\limits_{Y=y-\frac{1}{2}h}^{y+\frac{1}{2}h} \frac{1}{\sigma} \exp\left[-\frac{1}{2\sigma^2}(Y-y)^2\right] dY \\
&= \int\limits_{V=-\frac{h}{2\sigma}}^{\frac{h}{2\sigma}} \exp\left(-\frac{1}{2}V^2\right) dV.
\end{aligned}
$$

Since

$$
\int\limits_{V=-\frac{h}{2\sigma}}^{\frac{h}{2\sigma}} \exp\left(-\frac{1}{2}V^2\right) dV \underset{\sigma\to 0^+}{\longrightarrow} \int\limits_{-\infty}^{\infty} \exp\left(-\frac{1}{2}V^2\right) dV = \sqrt{2\pi},
$$

by dominated convergence, defining $L_{\max}\,(\sigma = 0, y) = \sqrt{2\pi}$ renders $L_{\max}$ a continuous function.

Thus the relative likelihood is

$$
R_{\max}(\sigma; y) = \frac{1}{\sqrt{2\pi}} \int\limits_{V=-\frac{h}{2\sigma}}^{\frac{h}{2\sigma}} \exp\left(\frac{1}{2}V^2\right) dV, \tag{3}
$$

and this is a continuous likelihood function with no singularity problems. The example may seem artificial, but a practical case is ungrouped mixtures of normal distributions. Here the density function has two singularities at each observation.

Example 3.2 A recent example is that of Berger et al. (1999). Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$ random variables while Y is independently $N(\lambda, \sigma_\theta^2)$, where $\sigma_\theta^2 = \exp(-n\theta^2) \leq 1$. The parameter of interest is θ and the nuisance parameter is λ. Berger et al. (1999) used this example to criticize the profile likelihood function in order to promote an integrated likelihood function. But because of the normal distribution of Y this is a slightly more general, although highly artificial, version of Example 3.1.

As in Example 3.1, when using the continuous approximation, since $\hat{\lambda}(\sigma_\theta) = y$, the single observation contributes the factor $1/\sigma_\theta$ to the likelihood function based on the X_i's. Thus the "profile" likelihood function of θ is proportional to

$$
\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]\frac{1}{\sigma_\theta} \propto \exp(n\bar{x}\theta) \to \infty \text{ as } \theta \to \pm\infty,
$$

depending on the sign of $\bar{x}$. This is referred to as a "very strange [profile] 'likelihood'" by Berger et al. (1999).

But, if one considers the finite precision of the measuring instrument, using (3) the actual profile likelihood function of θ is

$$L_{\max}(\theta; y) \propto \exp(n\bar{x}\theta)\sigma_\theta \int\limits_{V=-\frac{h}{2\sigma_\theta}}^{\frac{h}{2\sigma_\theta}} \exp\left(-\frac{1}{2}V^2\right) dV, \tag{4}$$

where $\sigma_\theta = \exp\left(-\frac{1}{2}n\theta^2\right) \leq 1$ (by assumption).

Berger et al. (1999) eliminate λ by integrating λ over the $N(\lambda, \sigma_\theta^2)$ density function of Y, to give the uniform integrated likelihood

$$L^U(\theta) \propto \exp\left[-\frac{1}{2}n\left(\bar{x}-\theta\right)^2\right] \propto \exp(n\bar{x}\theta)\sigma_\theta. \tag{5}$$

This is essentially assuming a uniform (improper) prior distribution of λ, $-\infty \leq \lambda \leq \infty$. The same result would be obtained by integrating out Y to give (5) as a marginal likelihood. For the case $n = 1, x = 1$, and $y = 0$, Fig. 1 shows some examples of how the relative profile likelihood functions of θ, which are (4) standardized to have a maximum of 1, are affected by changes in the precision h of the measuring instrument.

In fact, in this example the relative profile likelihood $L_{\max}$ in (4) includes the relative integrated or marginal likelihood (5) as a special case when h is large, indicating a lack of precision in measuring Y relative to the underlying variation of Y as determined by $\sigma_\theta \leq 1$. Figure 1 shows how $L_{\max}$ in (4) with $h = 5$ is practically indistinguishable from L^U in (5).

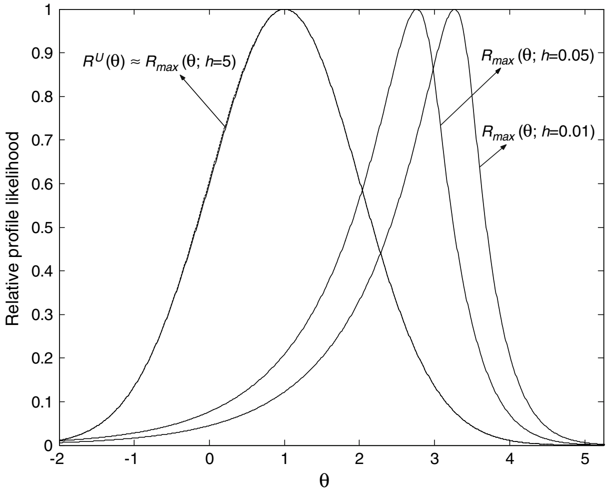

Fig. 1 Sensitivity of $R_{\max}(\theta; y)$ under changes in measuring precision in Example 3.2

Example 3.3 This example considers a threshold parameter α. A threshold parameter restricts the support of the corresponding random variable Y, and can be either a lower or an upper bound for Y. In contrast to the previous examples, the range of α in its corresponding likelihood function is determined by the observations and, consequently, will also be affected by their precision.

This situation occurs frequently in Reliability and Extreme Value Theory; for example, the three parameter Weibull and Fréchet distributions. The case discussed here is the three parameter lognormal distribution of Y that is given by

$$\log(Y-\alpha) = X(\alpha) \sim N(\theta,\sigma^2), \quad 0 \le \alpha < Y.$$

Then from a sample $y_1,\dots,y_n$ the restricted maximum likelihood estimates of θ and σ^2 for a given α are

$$\hat{\theta}(\alpha) = \bar{x}(\alpha) \quad \text{and} \quad \hat{\sigma}^2(\alpha) = \sum_{i=1}^{n} \left[x_i(\alpha) - \bar{x}(\alpha)\right]^2/n.$$

Using the continuous approximation, the resulting profile likelihood function of α is

$$\begin{aligned} L_{\max}(\alpha; y_1,\dots,y_n) &= L(\alpha,\hat{\theta}(\alpha),\hat{\sigma}(\alpha); y_1,\dots,y_n) \\ &\propto \left[\frac{1}{\hat{\sigma}^2(\alpha)}\right]^{\frac{n}{2}} \exp\left\{-\frac{1}{2\hat{\sigma}^2(\alpha)}\sum_{i=1}^{n}\left[x_i(\alpha)-\hat{\theta}(\alpha)\right]^2\right\}\prod_{i=1}^{n}\frac{1}{y_i-\alpha} \\ &\propto \left[\frac{1}{\hat{\sigma}^2(\alpha)}\right]^{\frac{n}{2}} \prod_{i=1}^{n}\frac{1}{y_i-\alpha}, \quad 0 \le \alpha < y_{(1)}, \end{aligned} \tag{6}$$

where $y_{(1)}$ is the smallest observation. The profile likelihood has a singularity at $\alpha = y_{(1)}$, inherited from the corresponding singularity of the joint density function.

As Meeker and Escobar (1998, p. 275) mention, there is a path in the parameter space for which the likelihood goes to infinity, in particular when $\sigma \to 0$ and $\alpha \to y_{(1)}$. It should be stressed that the likelihood approaches ∞ not necessarily because the probability of the data is large in that region of the parameter space, but instead because of a breakdown in the density approximation to the likelihood function. There is usually, though not necessarily always, a local maximum for the likelihood surface corresponding to the maximum of the correct likelihood based on the probability of the data shown in (2).

Therefore, one solution to this problem could be, as in the previous two examples, to use the correct likelihood contributions for each observation as in (2), implied by the finite precision of the measuring instrument. In this example however, a different and somewhat simpler approach can be used. The finite precision h of the measuring instrument is required to limit the range of α to be $\alpha < y_{(1)} - h$, Lawless (2003, p. 186), Barnard (1966). This would seem an equally

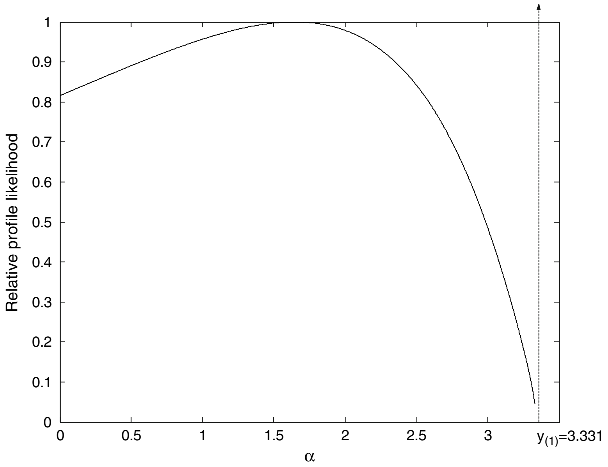

Fig. 2 Local versus global maximum of $R_{\max}(\alpha)$ in Example 3.3

reasonable way to incorporate the precision h, since it necessarily implies that the recorded $y_{(1)}$ could actually have been $y_{(1)} - h$, which would therefore be greater than α. That is, take $\alpha < y_{(1)} - h$ and use the continuous approximation to the likelihood function; the value of h depends on the precision of the measuring instrument.

To illustrate this other approach ten observations were simulated using $\alpha = 0$, $\theta = 2, \sigma = 1$ yielding in ascending order of magnitude

$$y = 3.331,\ 7.147,\ 7.661,\ 9.180,\ 9.253,\ 10.570,\\ 10.883,\ 14.006,\ 15.950,\ 51.400.$$

Assume that the precision of the measuring instrument is $h = 0.0005$. The resulting relative profile likelihood function of α is shown in Fig. 2. Note that $L_{\max}(\alpha)$ is not defined at $\alpha = y_{(1)}$ and $L_{\max}(\alpha)$ tends to infinity as it approaches $y_{(1)}$ from below. There is a local maximum $\hat{\alpha} = 1.65$, the maximum likelihood estimate, against which all the other values of (6) are standardized to give the relative profile likelihood function $R_{\max}(\alpha)$.

As shown on the graph there is a singularity at $\alpha = y_{(1)} = 3.331$, since $L_{\max}(\alpha) \to \infty$ as $\alpha \uparrow y_{(1)}$. Note that $R_{\max}(\alpha = y_{(1)} - h = 3.3305) = 0.0454$. Thus the singularity presents no practical problem for using the continuous approximation to the likelihood function. It is only necessary to bound α away from $y_{(1)}$ by h.

4 Discussion

The purpose of the preceding is to reemphasize that likelihood functions are proportional to probability functions and so cannot have singularities. Therefore criticisms of the profile likelihood that arise from singularities are invalid, and cannot be used to promote integrated or marginal likelihoods.

Acknowledgements The authors would like to thank a referee for useful comments that helped to clarify this manuscript. The first author would like to thank CIMAT for finantial support while working on the manuscript.

References

Barnard GA (1966) The use of the likelihood function in statistical practice. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, vol 1, pp 27–40

Barnard GA, Sprott DA (1983) Likelihood. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical science, vol 4. Wiley, New York, pp 639–644

Berger JO, Liseo B, Wolpert RL (1999) Integrated likelihood methods for eliminating nuisance parameters. Stat Sci 14:1–28

Edwards AWF (1992) Likelihood, expanded edition. The Johns Hopkins University Press, Baltimore

Kalbfleisch JG (1985) Probability and statistical inference. Springer, New York

Lawless JF (2003) Statistical models and methods for lifetime data. Wiley, Toronto

Lindsey JK (1998) Some statistical heresies (with discussion). Statistician 47:1–28

Meeker WQ, Escobar LA (1998) Statistical methods for reliability data. In: Wiley series in probability and statistics. Wiley, New York

Sprott DA (2000) Statistical inference in science. In: Springer series in statistics. Springer, New York