Metrika (2009) 69:265–282
DOI 10.1007/s00184-008-0217-0

Consistent likelihood-based estimation of a star-shaped distribution

Geurt Jongbloed

Published online: 4 December 2008

Abstract We consider the classical problem of nonparametrically estimating a star-shaped distribution, i.e., a distribution function F on $[0, \infty)$ with the property that $F(u)/u$ is nondecreasing on the set $\{u : F(u) < 1\}$. This problem is intriguing because of the fact that a well defined maximum likelihood estimator (MLE) exists, but this MLE is inconsistent. In this paper, we argue that the likelihood that is commonly used in this context is somewhat unnatural and propose another, so called 'smoothed likelihood'. However, also the resulting MLE turns out to be inconsistent. We show that more serious smoothing of the likelihood yields consistent estimators in this model.

Keywords Asymptotics · Censoring · Inverse problem · Maximum (smoothed) likelihood · Nonparametric · Shape constrained estimation

1 Introduction

The method of maximum likelihood goes back a long time. The name 'maximum likelihood' is usually credited to Fisher (1925), but opinions on who was the first to use the method differ (Le Cam 1990). Consider the context of estimating a distribution with density function f belonging to a class $\mathcal{F}$ of densities w.r.t. some dominating measure, based on an i.i.d. sample $X_1, \ldots, X_n$. Denote the ordered realized data points by $x_1 \leq x_2 \leq \cdots \leq x_n$. The MLE is then formally defined as the maximizer of the log likelihood function

$$\ell(f) = \sum_{i=1}^{n} \log f(x_i) \tag{1}$$

G. Jongbloed (✉)
Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands
e-mail: g.jongbloed@tudelft.nl

over the function class $\mathcal{F}$. In case the dominating measure is discrete, this quantity reflects the probability of observing $x_1, \ldots, x_n$ if f were the underlying density. The MLE can therefore be seen as the $f \in \mathcal{F}$ for which this probability is maximal. If the dominating measure is not discrete, there can be some trickiness in definition (1), due to the fact that densities are only defined up to sets of (dominating) measure zero. However, assuming the densities to have some continuity properties, (1) often works to define a proper MLE.

In the parametric context, where $\mathcal{F}$ is a class of densities that is smoothly parameterized by a parameter set $\Theta \subset \mathbb{R}^k$, there are results showing that under fairly general conditions the MLE is consistent and asymptotically normally distributed with rate of convergence $\sqrt{n}$ (see, e.g., van der Vaart 1998, Chapter 5). That assumptions are really needed, also in the parametric context, is clear from the following example of Quandt and Ramsey (1978), where ϕ denotes the standard normal density:

$$\mathcal{F} = \left\{ f_{\mu,\sigma} \,:\, (\mu, \sigma) \in \mathbb{R} \times (0, \infty) \right\}$$
$$\text{where} \quad f_{\mu,\sigma}(x) = \frac{1}{2}\phi(x) + \frac{1}{2\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}.$$

Indeed, taking $\mu = x_i$ for some i and $\sigma \downarrow 0$, shows that the corresponding log likelihood is unbounded; the MLE does not exist. In Ferguson (1982), a nice parametric example is given with parameter space $[0, 1]$, where the MLE is well defined but almost always converges to 1, no matter which parameter is used to generate the data.

In the nonparametric setting, the situation is more complicated, already starting with the definition of the likelihood function. In order to define a (nonparametric) MLE in a specific i.i.d. context, one usually starts off with the formal definition, where the likelihood is defined as the product of the density (w.r.t. some dominating σ-finite measure) evaluated at the observed data points. For example, if one considers the problem of estimating an increasing density (w.r.t. Lebesgue measure) on an interval $[0, a] \subset [0, 1]$ based on an i.i.d. sample, this definition makes sense and one can derive the maximum likelihood estimator (MLE) (related to the Grenander estimator, after Grenander 1956) as the nondecreasing density maximizing this objective function. This estimator is consistent. If one considers the problem of estimating a general distribution function on the real line, the formal definition of the likelihood does not make sense anymore, since not all these distributions have a density with respect to one single σ-finite dominating measure. Restricting the maximization to those distributions that do have a density w.r.t. Lebesgue measure, does not do the trick either. Using a basic kernel estimator with bandwidth tending to zero, shows that the likelihood is unbounded on this set and its maximizer is not well defined. The way out of this problem that is usually taken, is to leave the classical definition of the likelihood for continuous random variables and consider distributions that have a density w.r.t. counting measure on the observed data points (see Kiefer and Wolfowitz 1950). The log likelihood function is then defined as

$$\ell(F) = \sum_{i=1}^{n} \log P_F(X_i = x_i) = \sum_{i=1}^{n} \log(F(x_i) - F(x_i-)). \tag{2}$$

Maximizing this function over *all distribution functions* yields the empirical distribution function. This $\sqrt{n}$-consistent estimator has many desirable properties.

In spite of the fact that the maximum likelihood approach leads to good estimators in a variety of nonparametric problems, it can lead to ill-defined or inconsistent estimators. This point has been made by many authors. An example that is often given to stress this phenomenon, is that of estimating a star-shaped distribution. In that setting, one can define an MLE completely along the lines of that of a general distribution function, and this estimator is inconsistent. Noting that all convex distribution functions on $[0, a]$ are star-shaped and the previous discussion on the Grenander-type estimator, this might sound contradictory at first sight. The MLE is consistent over the class of increasing densities (small class) and that of all distribution functions (large class), but inconsistent over the class of star-shaped distributions (intermediate class).

In Sect. 2 we will follow the line of thought leading to an inconsistent MLE and review some more history of this problem, including references to consistent estimators. We also point out why the approach that leads to a successful MLE of a general distribution function, can be expected to fail when considering star-shaped distributions. In Sect. 3, we define an alternative likelihood function that, when restricted to the class of convex distribution functions, leads to an estimator related to the Grenander estimator and when considered as function over the class of all distribution functions, to the empirical distribution function. In fact, this definition unifies the likelihood approach in these three nested convex models, and is related to the so-called maximum smoothed likelihood estimator in the sense of Eggermont and LaRiccia (2001), where the level of smoothing is minimal. We derive an explicit representation of the maximizer of this (as well as a more heavily smoothed) likelihood over the class of star-shaped distributions.

In Sect. 4 we show that the more natural MLE defined in Sect. 3 (with minimal smoothing) is also inconsistent. However, we also show that the smoothed log likelihood function, where more smoothing is used, leads to a consistent estimator. In Sect. 5 we make some connections to other situations where MLE's are inconsistent.

2 Nonparametric estimation of a star-shaped distribution

In this paper we consider the problem of estimating a star-shaped distribution. Following Barlow et al. (1972), we define a distribution function F on $[0, \infty)$ to be star shaped if

$$x \mapsto \frac{F(x)}{x}$$

is increasing on the set $\{x \geq 0 : F(x) < 1\}$. The terminology is due to the fact that the set $\{(x, y) \in [0, \infty)^2 : 0 \leq F(x) < 1, y \geq F(x)\}$ is star-shaped as a subset of $\mathbb{R}^2$ in the sense that whenever (x, y) belongs to this set, the line segment connecting the origin with this point is also contained in this set.

The following missing data problem is a problem where exactly this shape constraint turns out to define the sampling distribution. Suppose Z has a distribution with

distribution function G, with $G(a) = 1$. Independent of Z, U is uniformly distributed on $(0, b)$ with $b \geq a$ and one observes $X = \max\{Z, U\}$. Then for $x \leq b$ the distribution function F of X is given by

$$F(x) = F_G(x) = P(X \leq x) = P(Z \leq x, U \leq x) = G(x)x/b \Rightarrow \frac{F(x)}{x} = \frac{G(x)}{b} \quad (3)$$

meaning that it is star-shaped. Moreover, to every star-shaped distribution function there corresponds a distribution function G in this way. Estimating a star-shaped distribution therefore corresponds to estimating an arbitrary distribution function G based on a sample from the star-shaped distribution function F_G (these distribution functions are in one-to-one correspondence). In this sense, the problem is that of estimating a sampling distribution in a *statistical inverse problem*. See Vardi (1989) for missing data models where the sampling distribution function turns out to be concave on $[0, \infty)$.

Saying that F is star-shaped geometrically means that for $y \in [0, F^{-1}(1)]$, the line connecting the points $(0, 0)$ and $(y, F(y))$ lies above the curve $x \mapsto F(x)$ on $[0, y]$ and below this curve on $[y, F^{-1}(1)]$. This shows in particular that convex distribution functions on $[0, a]$ are automatically star-shaped. In fact, the class of star-shaped distribution functions is a genuine superset of that of convex distribution functions, since also discontinuous (hence nonconvex) distribution functions like

$$F_u(x) = \begin{cases} 0 & \text{if } x \in [0, u) \\ x & \text{if } x \in [u, 1) \\ 1 & \text{if } x > 1 \end{cases} \quad (4)$$

for $u \in (0, 1)$ are star-shaped on $[0, 1]$. This also shows that a single σ-finite dominating measure for these distributions cannot be found. A restriction to absolutely continuous distribution functions also leads to difficulties, since one can construct star-shaped distribution functions that have an arbitrarily high derivative near the observed data points.

Analogously to the procedure leading to the empirical distribution function as MLE over the class of all distribution functions, Barlow et al. (1972) proceed by using log likelihood function (2) on the class of star-shaped distributions. Using representation (3), this yields

$$\begin{aligned} \ell(F) &= \sum_{i=1}^{n} \log(F(x_i) - F(x_i-)) \\ &= \sum_{i=1}^{n} \log(x_i(G(x_i) - G(x_i-))/x_n) \doteq \sum_{i=1}^{n} \log(G(x_i) - G(x_i-)), \end{aligned}$$

where $\doteq$ means equality up to an additive constant not depending on F. As a function of G, this criterion function is maximized by the discrete distribution function G giving mass $1/n$ to all observed data points x_i, where we suppose for the moment that there are no ties. Via (3) this again leads to the maximizer that assigns mass $x_i/(nx_n)$

to the point x_i. Writing $\mathbb{F}_n$ for the empirical distribution function of $x_1, \ldots, x_n$, the resulting estimator then becomes

$$\tilde{F}_n(x) = \mathbb{F}_n(x)\frac{x}{x_n} \wedge 1. \tag{5}$$

Using the law of large numbers and the fact that $X_{(n)} = \max_{1\le i\le n} X_i$ converges to $F^{-1}(1)$ a.s., it immediately follows that for any underlying star-shaped F,

$$\tilde{F}_n(x) \to \frac{xF(x)}{F^{-1}(1)} \wedge 1 \text{ a.s. for } n \to \infty.$$

This clearly implies inconsistency of this MLE. In Barlow et al. (1972), this result is derived only for the particular situation where F is the uniform distribution function on $(0, 1)$.

Let us elaborate on this inconsistency. The structure of the log likelihood function leads to an estimator that considers the data from the star-shaped distribution F as if these were generated directly by the underlying (unrestricted) distribution function G. The heart of the inconsistency problem can therefore be seen as the combination of a log in the function ℓ and the product relation between F and G, because by taking the log, the product structure defining the shape constraint on F disappears.

Before returning to likelihood based estimators, showing in fact that (2) is a very unnatural likelihood function to use in this context, a few words on alternative, consistent, nonparametric estimators for F. In Barlow et al. (1972), an isotonic estimator is defined via a least squares isotonic regression of a *naive* estimator for G, namely the piecewise constant function with value $\mathbb{F}_n(x_i)/x_i$ at the points x_i. They also propose another consistent isotonic estimator, the quantile estimator. In Wang (1988) another $\sqrt{n}$-consistent estimator of F is studied, the greatest star-shaped minorant of the empirical distribution function.

Let us now return to the procedure leading to the estimator (5). Using log likelihood function (2) is somewhat unnatural in this setting. The intersection of the class of star-shaped distributions and those with density w.r.t. counting measure on observed data points is *empty*. Indeed, if $x_1 < x_2$ and $0 < F(x_1) = F(x)$ for $x \in [x_1, x_2)$, $F(x)/x$ decreases on $[x_1, x_2)$. The procedure is nevertheless to optimize the likelihood over a class of discrete distribution functions, and then change the purely discrete distribution function to make up for the fact that it is not star-shaped. The mass that is introduced in between the observed data points during the second stage of this procedure, does not affect the log likelihood (2). In the next section a more natural MLE is defined, and characterized.

3 Smoothed likelihood-based estimation

As seen in the previous section, likelihood function (2) leads to an inconsistent MLE. A method of defining the likelihood function should take into account the fact that any star-shaped distribution *has* to assign mass to intervals in between successive observations (x_{i-1}, x_i), $i \ge 2$. Given observed data $0 < x_1 < x_2 < \cdots < x_n$, the following log likelihood function is a natural candidate:

$$\ell(F) = \sum_{i=1}^{n} \log(F(x_i) - F(x_{i-1})). \tag{6}$$

In the case there are ties in the data (note that this can happen in general, see, e.g., the distribution function on (4)), denote the distinct ordered observations by $y_1 < y_2 < \cdots < y_m$. Log likelihood (6) is then interpreted as

$$\ell(F) = \sum_{j=1}^{m} n_j \log(F(y_j) - F(y_{j-1})), \quad \text{where } n_j = |\{i : x_i = y_j\}|. \tag{7}$$

This type of log likelihood function was also considered under the name 'maximum product of spacings estimator' in the context of estimating a unimodal density in Shao (2001). Maximizing (6) over all distribution functions that are convex on their support, yields an estimator that is closely related to the Grenander estimator of a concave distribution function on $[0, \infty)$. Noting that one may restrict oneself to distribution functions having a left-continuous density, constant on intervals $(x_{i-1}, x_i]$, (6) can be written as

$$\ell(F) = \sum_{i=1}^{n} \log\left(f(x_i)(x_i - x_{i-1})\right) \doteq \sum_{i=1}^{n} \log f(x_i),$$

where again $\doteq$ denotes equality up to an additive constant not depending on f. Maximizing this function over the class of aforementioned densities is an instance of generalized isotonic regression, as described in Sect. 1.5 of Robertson et al. (1988). On $[0, x_n]$ it is the greatest convex minorant of the points $\{(x_i, \mathbb{F}_n(x_i)) : 0 \le i \le n\}$ whereas on $[x_n, \infty)$ it is one. Maximizing (7) over the larger class of all distribution functions on $[0, \infty)$, gives the empirical distribution function at the observed data points. Hence, (7) furnishes *one* single log likelihood function that can be maximized over classes of distribution functions to obtain an MLE over these classes.

Lemma 1 (Piecewise linear MLE) *In maximizing* (7) *over all star-shaped distribution functions, attention can be restricted to piecewise linear, continuous, distribution functions with knots* $y_1, \ldots, y_m$ *and* $F(y_m) = 1$.

Proof Consider an arbitrary star-shaped distribution function F on $[0, \infty)$. Based on this, we construct a star-shaped distribution function $\bar{F}$ of the type described in the statement of the lemma and show that $\ell(\bar{F}) \ge \ell(F)$. This shows that in maximizing (7) over all star-shaped distribution functions, we may restrict ourselves to distribution functions of that type.

Construct the piecewise linear distribution function $\bar{F}$ by connecting the points $(0, 0)$, $(y_j, F(y_j))$ for $1 \le j \le m - 1$ and $(y_m, 1)$. This distribution function is star-shaped. Indeed, that $j \mapsto \bar{F}(y_j)/y_j$ is nondecreasing is immediate. On the interval $(y_{j-1}, y_j]$, $1 \le j \le m - 1$, one can write

$$\bar{F}(y) = \frac{y_j F(y_{j-1}) - y_{j-1} F(y_j)}{y_j - y_{j-1}} + y \frac{F(y_j) - F(y_{j-1})}{y_j - y_{j-1}}.$$

The star-shape condition on F forces the first term to be nonpositive, since $y_j F(y_{j-1}) - y_{j-1}F(y_j) = y_j y_{j-1}(F(y_{j-1})/y_{j-1} - F(y_j)/y_j) \leq 0$. This implies that on $(y_{j-1}, y_j]$, $y \mapsto \bar{F}(y)/y$ is increasing. For $j = m$, this also holds since $F(y_m) \leq 1 = \bar{F}(y_m)$. The function $\bar{F}$ is a piecewise linear, continuous, distribution function with knots $y_1, \ldots, y_m$ and $F(y_m) = 1$. Moreover,

$$\ell(\bar{F}) - \ell(F) = n_m \left(\log(1 - F(y_{m-1})) - \log(F(y_m) - F(y_{m-1}))\right) \geq 0.$$

□

From now on we write $F_j = F(y_j)$ and identify the vector $(F_j)_{j=1}^m$ with the piecewise linear function having $F(y_j) = F_j$.

Theorem 1 *The piecewise linear continuous distribution function F with knots at $y_1, \ldots, y_m$ that maximizes (7) over all star-shaped distribution functions on $[0, \infty)$ is given by*

$$F(y_i) = T_n(y_i)\mathbb{F}_n(y_i), \quad \textit{where } T_n(y_i) = \prod_{j=i+1}^{m} \left(1 \wedge \frac{y_{j-1}\mathbb{F}_n(y_j)}{y_j \mathbb{F}_n(y_{j-1})}\right). \tag{8}$$

Proof Write $\lambda_i = F_{i-1}/F_i$. Then $F_i = \prod_{j=i+1}^m \lambda_j$ and in terms of this parametrization, computing the MLE boils down to maximizing

$$\begin{aligned}\phi(\lambda) &= \sum_{i=1}^{m} \left(n_i \log(1 - \lambda_i) + n_i \log \prod_{j=i+1}^{m} \lambda_j \right) \\ &= \sum_{i=1}^{m} \left(n_i \log(1 - \lambda_i) + \left(\sum_{j=1}^{i-1} n_j \right) \log \lambda_i \right)\end{aligned}$$

over the hypercube $\{\lambda = (\lambda_2, \ldots, \lambda_m) \in \mathbb{R}^{m-1} : \lambda_i \in [0, y_{i-1}/y_i], i = 2, 3, \ldots, m\}$. Differentiation with respect to λ_i yields that this function is maximized over λ_i by taking $\lambda_i = \left(1 - n_i / \sum_{j=1}^{i} n_j\right) \wedge (y_{i-1}/y_i)$, leading to

$$F_m = 1 \quad \text{and} \quad F_{i-1} = F_i \min \left\{ 1 - n_i / \sum_{1}^{i} n_k, y_{i-1}/y_i \right\} \quad \text{for } 2 \leq i \leq m.$$

Hence, $F(y_m) = 1$ and for $2 \leq i \leq m$

$$F(y_{i-1}) = \prod_{j=i}^{m} \min \left\{ 1 - n_j / \sum_{1}^{j} n_k, y_{j-1}/y_j \right\}.$$

Noting that

$$\prod_{j=i}^{m}\left(1-\frac{n_j}{\sum_{k=1}^{j} n_k}\right)=\prod_{j=i}^{m}\frac{\sum_{k=1}^{j-1} n_k}{\sum_{k=1}^{j} n_k}=\frac{1}{n}\sum_{k=1}^{i-1} n_k=\mathbb{F}_n(y_{i-1}),$$

gives

$$F(y_{i-1})=\mathbb{F}_n(y_{i-1})\prod_{j=i}^{m}\left(1\wedge\frac{y_{j-1}\sum_{k=1}^{j} n_k}{y_j\sum_{k=1}^{j-1} n_k}\right)=\mathbb{F}_n(y_{i-1})\prod_{j=i}^{m}\left(1\wedge\frac{y_{j-1}\mathbb{F}_n(y_j)}{y_j\mathbb{F}_n(y_{j-1})}\right).$$

□

The function maximizing (7) over all distribution functions, can be seen as a specific instance of a maximum smoothed likelihood estimator in the spirit of Eggermont and LaRiccia (2001). Indeed, taking $\tilde{F}_n$ as the linearly interpolated (and thus slightly smoothed) empirical distribution function, and denoting by f the piecewise constant density function corresponding to a piecewise linear star-shaped distribution function with knots $y_1, \ldots, y_m$, our log likelihood of F (apart from a constant not depending on F) is given by

$$\ell(F)=n\int \log f(x)\, d\tilde{F}_n(x).$$

In fact, defining a grid $0=z_0<z_1<\cdots<z_k=y_m$, we can in the same spirit define a maximum smoothed likelihood estimator based on the linearly interpolated empirical distribution function with knots in this grid. Copying the proof of Theorem 1, we get the following corollary.

Corollary 1 *The piecewise linear continuous distribution function F with knots at $0=z_0<z_1<\cdots<z_k$ maximizes*

$$F\mapsto\sum_{i=1}^{k} n_i\log(F(z_i)-F(z_{i-1}))\quad \textit{with}\ \ n_i=|\{j\,:\,x_j\in(z_{i-1},z_i]\}$$

over all star-shaped distribution functions of this type if and only if

$$F(z_j)=T_n(z_j)\mathbb{F}_n(z_j),\quad \textit{where}\ \ T_n(z_j)=\prod_{i=j+1}^{k}\left(1\wedge\frac{z_{i-1}\mathbb{F}_n(z_i)}{z_i\mathbb{F}_n(z_{i-1})}\right).\tag{9}$$

4 (In)consistency of the MSLE

We prove in this section that the maximizer of (7) over all star-shaped distributions is inconsistent in general. We do this by considering the case where F is the uniform distribution function on $[0, 1]$.

Theorem 2 *Let $X_1, X_2, \ldots$ be i.i.d. uniformly distributed random variables on* $[0, 1]$. *Then for each $x > 0$, the maximizer $\hat{F}_n$ of* (6) (*where x_i denotes the ith order statistic $X_{(i)}$ in the set $X_1, \ldots, X_n$*) *satisfies*

$$\hat{F}_n(x) \to^P x^{1/e} F(x) = x^{1/e+1}$$

as $n \to \infty$.

Proof Fix $x > 0$. Because F is continuous, there are no ties in the data. In the spirit of (8), define

$$T_n(x) = \prod_{\{i : X_{(i)} > x, X_{(i-1)}/X_{(i)} < 1-1/i\}} \frac{X_{(i-1)}/X_{(i)}}{1 - 1/i}.$$

Using the well known representation of uniform order statistics in terms of partial sums of i.i.d. exponential random variables $E_1, E_2, \ldots$, i.e.,

$$(X_{(1)}, X_{(2)}, \ldots, X_{(n)}) =^d \left(\frac{S_1}{S_{n+1}}, \frac{S_2}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}} \right), \quad \text{where } S_i = \sum_{j=1}^{i} E_j \tag{10}$$

we have

$$\begin{aligned} T_n(x) &=^d \prod_{\{i : S_i > x S_{n+1}\}} \left\{ \frac{i S_{i-1}/S_i}{i-1} \wedge 1 \right\} = \prod_{\{i : S_i > x S_{n+1}\}} \left\{ \left(1 + \frac{1 - i E_i/S_i}{i-1} \right) \wedge 1 \right\} \\ &= \exp\left(\sum_{\{i : S_i > x S_{n+1}\}} \log(1 + B_i) \right), \quad \text{with } B_i = 0 \wedge \frac{1 - i E_i/S_i}{i-1}. \end{aligned}$$

We will now show that

$$\left| \log T_n(x) - \frac{1}{e} \log x \right| \to^P 0 \quad \text{as } n \to \infty. \tag{11}$$

Define

$$\tilde{B}_i = 0 \wedge \frac{1 - E_i}{i-1}, \quad i = 2, 3, \ldots \tag{12}$$

and observe that by the triangle inequality

$$\begin{aligned} &\left| \sum_{i=\lceil nx \rceil}^{n} \log(1 + B_i) - \frac{1}{e} \log x \right| \le \left| \sum_{i=\lceil nx \rceil}^{n} (\log(1 + B_i) - B_i) \right| + \left| \sum_{i=\lceil nx \rceil}^{n} B_i - \tilde{B}_i \right| \\ &+ \left| \sum_{i=\lceil nx \rceil}^{n} \left(\tilde{B}_i - E\tilde{B}_i \right) \right| + \left| \sum_{i=\lceil nx \rceil}^{n} E\tilde{B}_i - \frac{1}{e} \log x \right| = |I_{1,n}| + |I_{2,n}| + |I_{3,n}| + |I_{4,n}|. \end{aligned}$$

Now, $I_{4,n} \to 0$ as $n \to \infty$, because

$$\sum_{i=\lceil nx \rceil}^{n} E\tilde{B}_i = \sum_{i=\lceil nx \rceil}^{n} \frac{\int_1^\infty (1-y)e^{-y}\,dy}{i-1} = -\sum_{i=\lceil nx \rceil}^{n} \frac{1/e}{i-1}$$
$$= -\frac{1}{e}\int_{nx}^{n} \frac{1}{y}\,dy - \left(\sum_{i=\lceil nx \rceil}^{n} \frac{1/e}{i-1} - \frac{1}{e}\int_{nx}^{n} \frac{1}{y}\,dy \right) = \frac{1}{e}\log x + R_n,$$

where $R_n \to 0$ as $n \to \infty$. Moreover, $I_{3,n} \to^P 0$ as $n \to \infty$ by Markov's inequality, since also

$$\mathrm{Var}\left(\sum_{i=\lceil nx \rceil}^{n} \tilde{B}_i \right) = \sum_{i=\lceil nx \rceil}^{n} \frac{\int_1^\infty (1-y)^2 e^{-y}\,dy - 1/e^2}{(i-1)^2} \to 0$$

as $n \to \infty$. Now consider $I_{2,n}$. By the strong law of large numbers,

$$R_m := \max_{j \geq m} \left| \frac{j}{S_j} - 1 \right| \to 0 \text{ almost surely as } m \to \infty. \tag{13}$$

Since $|B_i - \tilde{B}_i| \leq E_i |i/S_i - 1|/(i-1)$,

$$|I_{2,n}| \leq \sum_{i=\lceil nx \rceil}^{n} |B_i - \tilde{B}_i| \leq R_{\lceil nx \rceil} \sum_{i=\lceil nx \rceil}^{n} \frac{E_i}{i-1} \to^P 0$$

for $n \to \infty$ by (13) and Markov's inequality, because

$$E\left(\sum_{i=\lceil nx \rceil}^{n} \frac{E_i}{i-1} \right) = O(1) \quad \text{and} \quad \mathrm{Var}\left(\sum_{i=\lceil nx \rceil}^{n} \frac{E_i}{(i-1)} \right) \to 0.$$

For $I_{1,n}$, first note that for all $-1 < b \leq 0$,

$$0 \leq b - \log(1+b) = \sum_{j=2}^{\infty} \frac{(-b)^j}{j} = \frac{1}{2}b^2 \sum_{j=0}^{\infty} \frac{2}{j+2}(-b)^j \leq \frac{1}{2}b^2 \sum_{j=0}^{\infty} (-b)^j = \frac{b^2}{2(1+b)}.$$

Note that by Borel Cantelli and the strong law of large numbers

$$M_E := \sup_{j \geq 2} \frac{E_j}{\log j} < \infty \quad \text{and} \quad M_S := \max_{j \geq 1} \frac{j}{S_j} < \infty \text{ almost surely.} \tag{14}$$

Also using that $0 \le -B_i \le E_i(i/S_i)/(i-1) \le M_S E_i/(i-1)$, this leads to

$$0 \le I_{1,n} = \sum_{i=\lceil nx \rceil}^{n} (\log(1+B_i) - B_i) \le \sum_{i=\lceil nx \rceil}^{n} \frac{M_S^2 E_i^2}{2(i-1)^2} \cdot \left(1 - M_S \sup_{\lceil nx \rceil \le j \le n} \frac{E_j}{j-1}\right)^{-1}$$
$$\le \frac{1}{2} M_S^2 \sum_{i=\lceil nx \rceil}^{n} \frac{E_i^2}{(i-1)^2} \cdot \left(1 - M_S \frac{\log n}{nx-1} \sup_{\lceil nx \rceil \le j \le n} \frac{E_j}{\log j} \cdot \frac{(nx-1)\log j}{(j-1)\log n}\right)^{-1}$$
$$\le \frac{1}{2} M_S^2 \left(1 - M_S M_E \frac{\log n}{nx-1}\right)^{-1} \sum_{i=\lceil nx \rceil}^{n} \frac{E_i^2}{(i-1)^2} \to^P 0.$$

In the last step we use that the first factor is bounded in probability by (14) and the second, positive factor, converges to zero in expectation.

To see (11), note that for each $0 < \epsilon < x$, with probability tending to one,

$$\sum_{i=\lceil n(x-\epsilon) \rceil}^{n} \log(1+B_i) \le \log T_n(x) \le \sum_{i=\lceil n(x+\epsilon) \rceil}^{n} \log(1+B_i).$$

Finally, using (8) and the fact that the X_i's become dense in $[0,1]$ with probability one, $\hat{F}_n(x) \to^P x^{1/e} F(x)$ as $n \to \infty$. □

For the star-shaped distribution functions $F(x) = x^\alpha \wedge 1$, $\alpha \ge 1$, the estimator can be shown to converge to the function $F(x)x^{1/e^\alpha}$. This shows that the asymptotic bias factor depends on the underlying distribution function F for the MLE of Sect. 3, in contrast to the bias factor of the MLE of Sect. 2. The factor gets closer to one as the curvature of F increases. See Fig. 1 for an illustration of the inconsistency for $\alpha = 1$. For $\alpha = 2$, see Fig. 2, and note that the MLE of Theorem 1 is much closer to the true distribution function than in Fig. 1.

We now consider the maximum smoothed likelihood estimator of Corollary 1, under some assumptions on the distance between successive points z_i of the grid and the underlying star-shaped distribution. Theorem 3 shows, without using the explicit representation of the estimator, that it is consistent under a 'strict star-shaped' condition and a condition on the level of smoothing of the empirical distribution function.

Theorem 3 *Suppose that the underlying star-shaped distribution function F is strictly star-shaped in the following sense. For some $t > 0$, there exist $c, \beta, \delta > 0$ such that for $t \le z < F^{-1}(1)$*

$$\frac{F(z+h)}{z+h} - \frac{F(z)}{z} \ge c h^\beta \quad \text{for all } 0 < h < \delta \wedge (F^{-1}(1) - z). \tag{15}$$

Moreover, suppose that the grid points z_i are taken in such a way that for all i

$$z_i - z_{i-1} \ge \kappa_n n^{-1/(2\beta)} \quad \text{with } 0 < \kappa_n \to \infty \text{ as } n \to \infty.$$

Then, for all $z > t$, $\hat{F}_n(z) \to^P F(z)$ as $n \to \infty$.

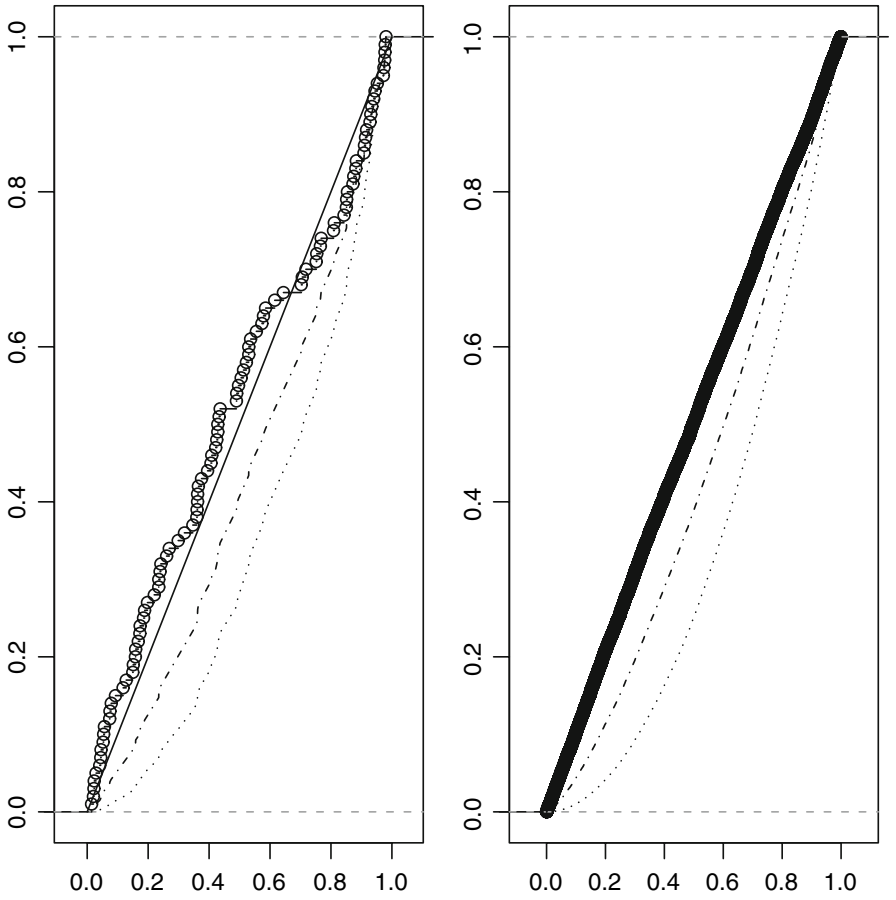

Fig. 1 The *left* picture shows the empirical distribution function, the estimator (5) (*dotted*) and the estimator of Theorem 1 (*dashed*) based on a sample of size $n = 100$ from the uniform distribution on (0, 1). The *solid line* is the underlying uniform distribution function. The *right* picture shows these estimators based on a sample of size $n = 10.000$ from the same distribution

Proof Fix $x > t$. By Donsker's theorem we have that $V_n = \sqrt{n}\|\mathbb{F}_n - F\|_\infty = O_P(1)$. Moreover, a.s., for n sufficiently large, $F(x)/2 < \mathbb{F}_n(x) < 2F(x)$. Hence, for $z_{j-1} \geq x$,

$$\frac{z_{j-1}\mathbb{F}_n(z_j)}{z_j\mathbb{F}_n(z_{j-1})}$$

$$= \frac{z_{j-1}(\mathbb{F}_n(z_j) - F(z_j))}{z_j\mathbb{F}_n(z_{j-1})} + \frac{z_{j-1}F(z_j)}{z_j}\left(\frac{1}{\mathbb{F}_n(z_{j-1})} - \frac{1}{F(z_{j-1})}\right) + \frac{F(z_j)z_{j-1}}{F(z_{j-1})z_j}$$

$$\geq -\frac{2z_m V_n}{\sqrt{n}xF(x)} - \frac{4z_m V_n}{\sqrt{n}xF(x)^2} + \frac{F(z_j)z_{j-1}}{F(z_{j-1})z_j}$$

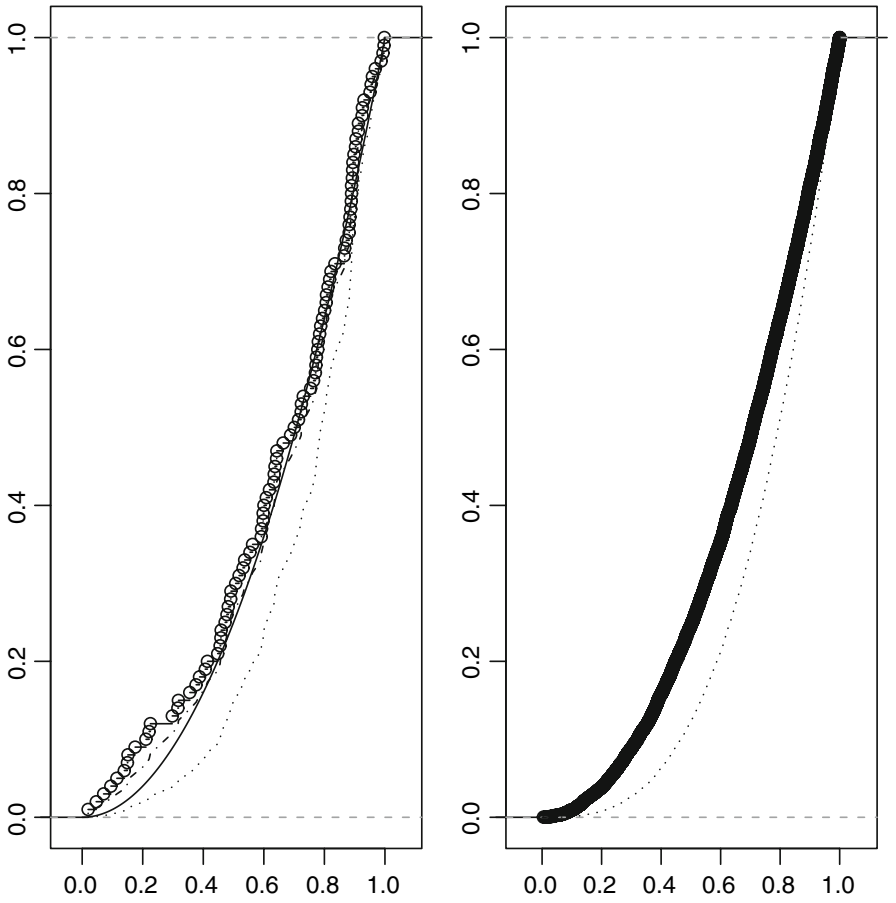

Fig. 2 The *left* picture shows the empirical distribution function, the estimator (5) (*dotted*) and the estimator of Theorem 1 (*dashed*) based on a sample of size $n = 100$ from the distribution function $F(x) = x^2$ on (0, 1). The *solid line* is the underlying distribution function. The *right* picture shows these estimators based on a sample of size $n = 10.000$ from the same distribution

$$\geq -\frac{2z_m V_n}{\sqrt{n}xF(x)} - \frac{4z_m V_n}{\sqrt{n}xF(x)^2} + 1 + \frac{z_{j-1}}{F(z_{j-1})}\left(\frac{F(z_j)}{z_j} - \frac{F(z_{j-1})}{z_{j-1}}\right)$$

$$\geq 1 - \lambda\frac{V_n}{\sqrt{n}} + xc(z_{j+1} - z_j)^\beta \geq 1 + n^{-1/2}\left(xc\kappa_n^\beta - \lambda V_n\right).$$

Note that with probability tending to one this right hand side will be bigger than one uniformly in j such that $z_{j-1} > x$. This implies that with probability tending to one, $T_n(z_j) = 1$, where T_n is as defined in (9). This proves the lemma. □

Theorem 3 can, e.g., be applied to star-shaped distributions with $F(x) = xG(x)$ where G is a distribution function with density $g(x) = G'(x) \geq c > 0$ on $[0, F^{-1}(1)]$. Then (15) holds with $\beta = 1$ and taking the grid such that $\max_i(z_i - z_{i-1}) \to 0$ and

$\sqrt{n}\min_i(z_i - z_{i-1}) \to \infty$, yields a consistent estimator. Theorem 3 cannot be applied to star-shaped functions that have linear parts on which $x \mapsto F(x)/x$ is constant, such as the uniform distribution. However, also for that type of functions, consistency can be proved. We consider the estimator that is obtained by taking $z_i = X_{(ik_n)}$ where k_n is a sequence that tends to infinity at a slower rate than n. In fact this corresponds to a random grid, jumping through the set of observed data points such that the jump size tends to zero, but the number of observations jumped over at each step tends to infinity. This density estimate is related to that used in Tsybakov and Van der Meulen (1996) to estimate the entropy of a probability density. Also other choices for the grid could be chosen, but we do not pursue that here.

Theorem 4 *Let F be a star-shaped distribution function with support* $[0, 1]$. *Let $k = k_n$ be a sequence of positive numbers, tending to infinity at rate $o(n)$. Let $\hat{F}_n$ be the estimator defined in Corollary 1, with $z_j = X_{(jk_n)}$. Then for each $x > 0$, $\hat{F}_n(x) \to^P F(x)$.*

Proof Denote by Q the quantile function F^{-1} belonging to F, and note that for all $0 < u < v < 1$, the star-shaped condition on F implies that

$$Q(u) \geq u \quad \text{and} \quad Q(v) - Q(u) \leq \frac{Q(u)}{u}(v-u)$$
$$\Rightarrow \frac{Q(v) - Q(u)}{Q(v)} \leq \frac{Q(u)(v-u)}{Q(v)u} \leq \frac{v-u}{u}.$$

Note that the latter inequality also trivially holds if $u = v$. By the representation of uniform order statistics on $(0, 1)$ in terms of i.i.d. exponential random variables given in (10) and the quantile method,

$$(X_{(1)}, \ldots, X_{(n)}) =^d \left(Q\left(\frac{S_1}{S_{n+1}}\right), Q\left(\frac{S_2}{S_{n+1}}\right), \ldots, Q\left(\frac{S_n}{S_{n+1}}\right)\right).$$

Now, also using the inequality derived above,

$$\begin{aligned}
\frac{\mathbb{F}_n(Z_{(j)})Z_{(j-1)}}{\mathbb{F}_n(Z_{(j-1)})Z_{(j)}} &= \frac{\mathbb{F}_n(X_{(k_n j)})X_{(k_n(j-1))}}{\mathbb{F}_n(X_{(k_n(j-1))})X_{(k_n j)}} =^d \frac{jQ(S_{k_n(j-1)}/S_{n+1})}{(j-1)Q(S_{k_n j}/S_{n+1})} \\
&= \frac{(j-1)Q(S_{k_n j}/S_{n+1}) + Q(S_{k_n j}/S_{n+1}) + j(Q(S_{k_n(j-1)}/S_{n+1}) - Q(S_{k_n j}/S_{n+1}))}{(j-1)Q(S_{k_n j}/S_{n+1})} \\
&= 1 + \frac{1 - j(Q(S_{k_n j}/S_{n+1}) - Q(S_{k_n(j-1)}/S_{n+1}))/Q(S_{k_n j}/S_{n+1})}{j-1} \\
&\geq 1 + \frac{1 - j(S_{k_n j} - S_{k_n(j-1)})/S_{k_n(j-1)}}{j-1}.
\end{aligned}$$

Fix $x > 0$ and note that

$$T_n(x) = \prod_{\{j\,:\,Z_j > x\}} \frac{\mathbb{F}_n(Z_{(j)})Z_{(j-1)}}{\mathbb{F}_n(Z_{(j-1)})Z_{(j)}} \wedge 1 \geq^d \exp\left(\sum_{\{j\,:\,Q(S_{jk_n}/S_{n+1}) > x\}} \log(1 + C_j) \right)$$

with

$$C_j = 0 \wedge \frac{1 - j(S_{k_n j} - S_{k_n(j-1)})/S_{k_n(j-1)}}{j-1}.$$

Arguing as in the proof of Theorem 2, observe that for j large, $S_{(j-1)k_n} \approx (j-1)k_n$, giving

$$\sum_{\{j\,:\,Q(S_{jk_n}/S_{n+1})>x\}} \log(1+C_j) \approx \sum_{j=\lceil nx/k_n\rceil}^{\lceil n/k_n\rceil} C_j \approx \sum_{j=\lceil nx/k_n\rceil}^{\lceil n/k_n\rceil} \tilde{C}_j, \tag{16}$$

where

$$\tilde{C}_j = 0 \wedge \frac{1-(S_{k_n j} - S_{k_n(j-1)})/k_n}{j-1} = 0 \wedge \frac{1 - k_n^{-1}\sum_{\ell=k_n(j-1)+1}^{k_n j} E_\ell}{j-1}$$
$$= -\frac{G_{k_n}^{(j)} \vee k_n - k_n}{k_n(j-1)}$$

with $G_{k_n}^{(j)} \sim \text{Gamma}(k_n)$. Note the crucial difference of this expression with (12), where a single E_i appears in contrast to the mean of an increasing number of such exponentials. Now, denoting the Gamma(p) density by ϕ_p, we get using integration by parts

$$E\left(G_{k_n}^{(j)} \vee k_n - k_n\right) = \int_{k_n}^{\infty} (y-k_n)\phi_{k_n}(y)\,dy$$
$$= \frac{1}{(k_n-1)!}\int_{k_n}^{\infty}(y^k - k y^{k-1})e^{-y}\,dy = \frac{k_n^{k_n}e^{-k_n}}{(k_n-1)!}.$$

Also using the Stirling approximation $n! \sim \sqrt{2\pi}n^{n+1/2}e^{-n}$ as $n \to \infty$ (in the sense that the ratio tends to one) and the logarithmic approximation to the sum of $1/(j-1)$ we get

$$E\left|\sum_{j=\lceil nx/k_n\rceil}^{\lceil n/k_n\rceil}\tilde{C}_j\right| \le \sum_{j=\lceil nx/k_n\rceil}^{\lceil n/k_n\rceil} E|\tilde{C}_j| = \sum_{j=\lceil nx/k_n\rceil}^{\lceil n/k_n\rceil} \frac{E\left(G_{k_n}^{(j)} \vee k_n - k_n\right)}{k_n(j-1)} \sim \frac{-\log x}{\sqrt{2\pi k_n}} \to 0$$

as $n \to \infty$. Since also the variance of this sum tends to zero, $T_n(x) \to^P 1$, implying that $\hat{F}_n(y) - \mathbb{F}_n(y) \to^P 0$ for all $y > x$. The steps in (16) can be made rigorous in the same way as was done in the proof of Theorem 2. □

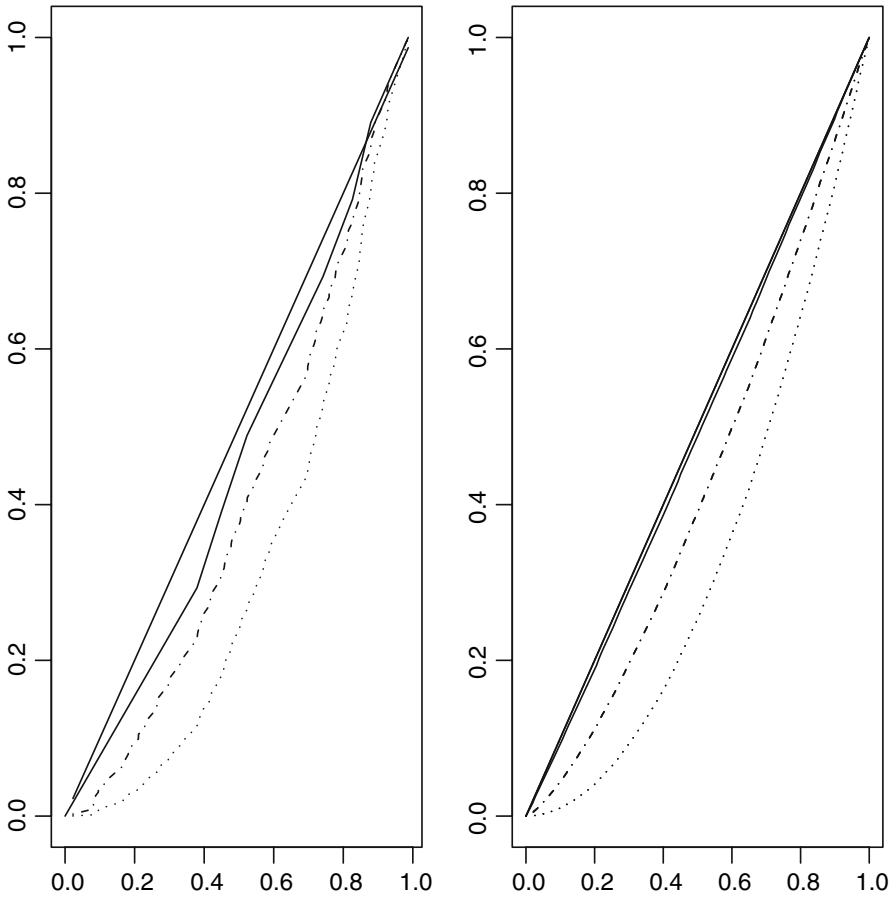

Fig. 3 The *left* picture shows the estimator (5) (*dotted*), the estimator of Theorem 1 (*dashed*) and the estimator of Corollary 1 (*solid*) with $z_i = x_{10i}$ based on a sample of size $n = 100$ from the uniform distribution (0, 1). The *straight solid line* is the underlying uniform distribution function. The *right* picture shows these estimators based on a sample of size $n = 10.000$ from the same distribution, with $z_i = x_{100i}$

Figure 3 shows three estimators based on samples of sizes $n = 100$ and $n = 10.000$ from the uniform distribution on (0, 1): the estimator of (5), Theorem 1 and the consistent maximum smoothed likelihood estimator of Corollary 1.

5 Discussion

An intriguing example of a problem where the (nonparametric) MLE is inconsistent, is that of estimating a star-shaped distribution. We argue that the likelihood function that is usually considered in that setting, is somewhat unnatural. It corresponds to the product of point masses at observation points, where actually a star-shaped distribution has to assign quite some mass to points outside the set of observations. We propose another likelihood that takes these masses into account. It can also be used

in the context estimating a convex distribution function on an interval $[0, b]$ as well as in the context of estimating a general distribution function (without any shape constraints). In these settings, it leads to the usual MLEs: a Grenander-type estimator and the empirical distribution function respectively. This alternative likelihood function can be interpreted as a smoothed likelihood in the spirit of Eggermont and LaRiccia (2001), where the level of smoothing is minimal. We show this alternative MLE is also inconsistent. In view of the observation that the objective function is exactly the same as in the setting of estimating a convex distribution function and a general distribution function, this is an interesting result. This objective function optimized over all convex distribution functions leads to a consistent estimator, as well as optimized over all distribution functions. But optimizing the same function over the intermediate class of star-shaped distribution functions, leads to an inconsistent estimator.

We finally show that increasing the level of smoothing in the smoothed likelihood, leads to a consistent likelihood based method. We expect this result also to hold true in other situations where the MLE is not consistent. For example the problem where the uniform distribution in (3) is replaced by a more general distribution or the problem of estimating an increasing failure rate average (IFRA) distribution as considered in Boyles et al. (1985). Another example is work in progress (Groeneboom et al. 2008). There the inconsistency problem, addressed in Maathuis and Wellner (2008) in the context of estimating a bivariate distribution function in the current status problem with continuous marks, is handled by using a smoothed version of the empirical distribution of the observed data instead of the empirical distribution function in the definition of the log likelihood.

Acknowledgments Thanks to two referees, whose comments lead to various improvements in the text and streamlining of the proof of Theorem 2.

References

Barlow RE, Bartholomew DJ, Bremner JN, Brunk HD (1972) Statistical inference under order restrictions. Wiley, New York

Boyles RA, Marshall AW, Proshan F (1985) Inconsistency of the maximum likelihood estimator of a distribution having increasing failure rate average. Ann Stat 13:413–417

Eggermont PPB, LaRiccia V (2001) Maximum penalized likelihood estimation volume I: density estimation. Springer, New York

Ferguson TS (1982) An inconsistent maximum likelihood estimate. J Am Stat Assoc 77:831–834

Fisher RA (1925) Theory of statistical estimation. Proc Camb Phil Soc 22:700–725

Grenander U (1956) On the theory of mortality measurement, part II. Skandinavisk Aktuarietidskrift 39:125–153

Groeneboom P, Jongbloed G, Witte BI (2008) Maximum smoothed likelihood estimation of a bivariate distribution function of interval censored survival times and continuous marks. Work in progress

Kiefer J, Wolfowitz J (1950) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Stat 27:887–900

Le Cam L (1990) Maximum Likelihood: an introduction. Int Stat Rev 58:153–171

Maathuis MH, Wellner JA (2008) Inconsistency of the MLE of interval censored survival times and continuous marks. Scand J Stat 35:83–103

Quandt RE, Ramsey JL (1978) Estimating mixtures of normal distributions and switching regressions. J Am Stat Assoc 73:730–738

Robertson T, Wright FT, Dykstra RL (1988) Order restricted statistical inference. Wiley, Chichester

Shao Y (2001) Consistency of the maximum product of spacings method and estimation of a unimodal distribution. Stat Sin 11:1125–1140

Tsybakov AB, Van der Meulen EC (1996) Root-n consistent estimators of entropy for densities with unbounded support. Scand J Stat 23:75–83

van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge

Vardi Y (1989) Multiplicative censoring, renewal processes, deconvolution and decreasing density: non-parametric estimation. Biometrika 17:84–99

Wang JL (1988) Optimal estimations of starshaped distribution functions. Stat Decis 6:21–32