ELSEVIER

Statistics & Probability Letters 45 (1999) 41–47

STATISTICS & PROBABILITY LETTERS

www.elsevier.nl/locate/stapro

Isotonic estimation for grouped data

Michael Woodroofe *,[1], Rong Zhang[1]

Statistics Department, University of Michigan, Ann Arbor MI 48109, USA

Received October 1998; received in revised form January 1999

Abstract

A non-parametric estimator of a non-increasing density is found in a class of piecewise linear functions when the data consist only of counts. An EM-Algorithm for computing the estimator is developed, and the iterates in the algorithm are shown to converge to the maximum likelihood estimator. Potential applications to distance sampling models are described and illustrated with a numerical example. © 1999 Elsevier Science B.V. All rights reserved

Keywords: Counts; Distance sampling; EM-Algorithm; Maximum likelihood estimation

1. Introduction

The problem considered here is estimating a non-increasing density from grouped (or binned) data. Motivation for this project derives, in part, from distance sampling models, as described by Buckland et al. (1992). There an observer searches for hidden objects and records the distance X from the object to his/her closest point of approach. Let $g(x)$ be the probability of finding an object x distance away. The conditional density function of X given an object is detected is $f(x)=g(x)/\mu$, where $\mu=\int_0^\infty g(x)\,\mathrm{d}x$ is the unconditional probability of finding an object. Grouped data are common in such studies, and there is interest in the density of X, especially at $x=0$. It is reasonable to suppose that the density of X (given observation) is non-increasing here and, therefore, natural to consider isotonic methods, as described by Robertson et al. (1989). Isotonic methods were used by Johnson and Routledge (1985), but do not appear to have been exploited in much detail for distance sampling. It is hoped that this paper may help form a connection between distance sampling and isotonic methods and provide a non-parametric analysis to complement the parametric analyses in Buckland et al. (1992). Further motivation is provided by Bickel and Fan (1996), who suggest grouping data before applying isotonic methods in the context of unimodal density estimation.

In the next section a non-parametric maximum likelihood estimator (hereafter NPMLE) for f is sought in the class of non-increasing, continuous piecewise linear densities with knots at the boundaries of the class intervals. Characterization of the NPMLE and an iterative computational scheme are presented and illustrated by a numerical example. The resulting iteration is an EM-Algorithm and this fact is useful in the proof of

* Corresponding author.

[1] Research supported by the National Science Foundation under DMS9626347.

PII: S0167-7152(99)00039-5

convergence. The latter is presented in Section 3 and may be of technical interest because it provides an example in which the EM-Algorithm converges to the maximum likelihood estimator when the latter is a boundary point.

2. The NPMLE

2.1. The likelihood function

Let $X_1,\ldots,X_n \sim^{\text{ind}} f$ be independent and identically distributed non-negative random variables with an unknown density f, but suppose that f is known to be non-increasing on $[0,\infty)$. Let $0=x_0<\cdots<x_m<\infty$ and suppose that only the counts

$$n_k=\#\{i\colon x_{k-1}<X_i\leqslant x_k\},\quad k=1,\ldots,m$$

are observed. Here $x_1,\ldots,x_m$ are regarded as constants that are imposed externally. For example, if distances are only recorded to the nearest meter, then $x_k=k$. With such data, an estimate of f is sought within the class of continuous, non-increasing, piecewise linear densities on $[0,x_m]$ with knots at $x_0,\ldots,x_m$. Let $\mathscr{F}_1$ denote the class of such densities f, and let $p_k=P_f\{x_{k-1}<X_1\leqslant x_k\}$. Then $p_k=(f_{k-1}+f_k)\Delta x_k/2$, where $f_k=f(x_k)$ and $\Delta x_k=x_k-x_{k-1}$. The log-likelihood function is

$$l(f)=\sum_{k=1}^{m} n_k\log(p_k)=\sum_{k=1}^{m} n_k\log(f_{k-1}+f_k)+C, \tag{1}$$

where C does not depend on f. This is to be maximized subject to the constraints

$$f_0\geqslant f_1\geqslant\cdots\geqslant f_m\geqslant 0 \tag{2}$$

and $\sum_{k=1}^{m} p_k=1$. Observe that $\sum_{k=1}^{m} p_k=\sum_{k=0}^{m} f_k\Delta y_k$, where $y_{-1}=0$ and $y_k=(x_k+x_{k+1})/2$ for $k=0,\ldots,m$, and $x_{m+1}=x_m$. Thus, the second constraint may be rewritten as

$$\sum_{k=0}^{m} f_k\Delta y_k=1. \tag{3}$$

It is convenient to use the same symbol f to denote both the function and the vector $f=(f_0,\ldots,f_m)$. Let Ω be the set of $f=(f_0,\ldots,f_m)$ for which (2) holds, and let Ω_0 be the subset on which (3) holds. In the remainder of the paper, it is required that

$$n_k>0,\quad k=1,\ldots,m. \tag{4}$$

Then l attains its maximum on Ω_0 at a unique f, say $f=\hat{f}$.

2.2. A related problem

Likelihood function (1) arises in a related problem. Let $\mathscr{F}_0$ denote the class of non-increasing, left-continuous, piecewise constant densities with knots at y_k, $k=-1,\ldots,m$. Thus, each $f\in\mathscr{F}_0$ may be written as $f(x)=f_k$ for $y_{k-1}<x\leqslant y_k$ for $k=0,\ldots,m$, where again $f_k=f(x_k)$. Then l is the log-likelihood function for the related problem, and (3) is the condition that $\int_0^{y_m} f(x)\,\mathrm{d}x=1$. So, it suffices to solve the related problem. The related problem would have a simple solution, if there were additional data. Let

$$r_{0k}=\#\{i\colon x_k<X_i\leqslant y_k\},$$

$$r_{1k}=\#\{i\colon y_{k-1}<X_i\leqslant x_k\}$$

for $k=0,\ldots,m$. Thus, $r_{0m}=r_{10}=0$. Next, let l^* denote the log-likelihood function for the problem in which $r_{00},\ldots,r_{1m}$ are observed:

$$l^*(f)=\sum_{k=0}^{m}(r_{0k}+r_{1k})\log(f_k)+C^*,$$

where C^* does not depend on f. Then l^* is maximized subject to (2) and (3) by

$$f_k^*=\min_{-1\leqslant i<k}\ \max_{k\leqslant j\leqslant m}\frac{G(y_j)-G(y_i)}{y_j-y_i},$$

where $G(0)=0$ and

$$G(y_k)=\frac{1}{n}\sum_{j=0}^{k}(r_{0j}+r_{1j})$$

for $k=0,\ldots,m$. Alternatively, letting F^* be a continuous piecewise linear function for which $F^*(0)=0$ and $F^*(y_k)=f_0^*\Delta y_0+\cdots+f_k^*\Delta y_k$ for $k=0,\ldots,m$, F^* is the least concave majorant of G. These assertions follow easily from Example 1.5.7 of Robertson et al. (1989).

2.3. An EM algorithm

For the case in which only $n_1,\ldots,n_m$ are observed, the maximum likelihood estimator may be found from the EM-Algorithm. First observe that

$$r_{0,k-1}+r_{1k}=n_k \tag{5}$$

for $k=1,\ldots,m$. Let E_f^* and P_f^* denote conditional expectation and probability given $n_1,\ldots,n_m$, when f is the density. When f is a piecewise constant density, as described above, it is easily seen that $r_{00},\ldots,r_{0,m-1}$ are conditionally independent binomial random variables, such that

$$E_f^*(r_{0k})=\left(\frac{f_k}{f_k+f_{k+1}}\right)n_{k+1}=r_{0k}^f \quad \text{say,} \tag{6}$$

for $k=0,\ldots,m-1$ and, therefore, that $E_f^*(r_{1k})=f_kn_k/(f_{k-1}+f_k)=r_{1k}^f$, say, for $k=1,\ldots,m$. The E-Step in the EM Algorithm is to compute $Q(f;g)=E_f^*[l^*(g)]$ for $f,g\in\mathscr{F}_0$. This is easy using (5) and (6), since l^* is linear in $r_{00},\ldots,r_{1m}$, and

$$Q(f;g)=\sum_{k=0}^{m}(r_{0k}^f+r_{1k}^f)\log(g_k)+C_f,$$

where $r_{10}^f=0=r_{0m}^f$ and C_f does not depend on g. The M-Step is to maximize $Q(f;g)$ with respect to $g\in\mathscr{F}_0$. This too is easy, since $Q(f;g)$ has the same form as $l^*(g)$; $Q(f;g)$ is maximized by

$$g_k^*=\min_{-1\leqslant i<k}\ \max_{k\leqslant j\leqslant m}\frac{G_f(y_j)-G_f(y_i)}{y_j-y_i},$$

where G_f is a continuous piecewise linear function with knots at $y_{-1},y_0,\ldots,y_m$ and values $G_f(0)=0$, $G_f(y_m)=1$, and $G_f(y_k)=(1/n)\sum_{j=0}^{k}(r_{0j}^f+r_{1j}^f)$ for $k=0,\ldots,m-1$. By (5) and (6), there is the alternative expression for G_f :

$$G_f(y_k)=\frac{1}{n}\left[\sum_{j=1}^{k}n_j+\left(\frac{f_k}{f_k+f_{k+1}}\right)n_{k+1}\right]. \tag{7}$$

Write $g^* = M(f)$. Then the EM-Algorithm takes the simple form: starting with any $f^0 \in \mathscr{F}_0$, let

$$f^q = M(f^{q-1}) \tag{7'}$$

for $q = 1, 2, \ldots$. In Section 3, the iterates f^q are shown to converge to $\hat{f}$ as $q \to \infty$ for any f^0 for which $f^0_m > 0$. So, the NPMLE satisfies the relation

$$\hat{f}_k = \min_{-1 \leqslant i < k} \max_{k \leqslant j \leqslant m} \frac{G_{\hat{f}}(y_j) - G_{\hat{f}}(y_i)}{y_j - y_i} \tag{8}$$

for $k = 0, \ldots, m$.

Relation (8) suggests a simpler, approximate NPMLE. If ratio on the right-hand side of (7) is replaced by $\frac{1}{2}$, then $G(y_k)$ is replaced by $\overline{G}(y_k) = (n_1 + \cdots + n_k + 0.5 \times n_{k+1})/n$ for $k = 0, \ldots, m-1$, and $\hat{f}_k$ by

$$\bar{f}_k = \min_{-1 \leqslant i < k} \max_{k \leqslant j \leqslant m} \frac{\overline{G}(y_j) - \overline{G}(y_i)}{y_j - y_i} \tag{8'}$$

for $k = 0, \ldots, m$, where $\overline{G}(0) = 0$ and $\overline{G}(y_m) = 1$. Relation (8′) does not require iteration, and $\bar{f}$ is called the *approximate NPMLE* below.

Example. *The Wooden Stake Data*. Buckland et al. (1992), report data in which X is the distance from a hidden object to the observer's closest point of approach, and there are 20 class intervals of length 1 m each. The estimates $\hat{f}_k$ and $\bar{f}_k$ for this data set are shown in Table 1. Bootstrap estimates of $E(\hat{f}_k)$, $E(\bar{f}_k)$ and the standard deviations are included in Table 1 too. Table 1 also includes a parametric MLE $\tilde{f}_k$, which is

Table 1
The wooden stake data

x_k	n_k	$\hat{f}_k$	Mean $(\hat{f}^*_k/\hat{f}_k)$	*se* $(\hat{f}^*_k)$	$\bar{f}_k$	Mean $(\bar{f}^*_k/\bar{f}_k)$	*se* $(\bar{f}^*_k)$	$\tilde{f}$
0	–	0.1543	0.9723	0.0270	0.1293	0.9440	0.0115	0.1204
1	83	0.1043	1.0621	0.0121	0.1121	0.9440	0.0078	0.1179
2	61	0.1043	1.0014	0.0095	0.1044	1.0059	0.0075	0.1107
3	73	0.1043	0.9283	0.0102	0.1005	0.9375	0.0072	0.0998
4	56	0.0709	1.0959	0.0087	0.0727	1.1038	0.0064	0.0866
5	31	0.0709	1.0154	0.0064	0.0727	1.0104	0.0059	0.0729
6	59	0.0709	0.9889	0.0062	0.0727	0.9616	0.0058	0.0601
7	44	0.0709	0.9277	0.0080	0.0654	0.9513	0.0061	0.0494
8	40	0.0409	1.0340	0.0093	0.0467	1.0094	0.0058	0.0416
9	20	0.0295	1.1044	0.0042	0.0297	1.1759	0.0041	0.0366
10	11	0.0295	1.0550	0.0031	0.0297	1.0701	0.0031	0.0341
11	29	0.0295	1.0310	0.0028	0.0297	1.0392	0.0028	0.0331
12	13	0.0295	1.0109	0.0026	0.0297	1.0149	0.0026	0.0327
13	19	0.0295	0.9933	0.0026	0.0297	0.9936	0.0026	0.0319
14	16	0.0295	0.9751	0.0027	0.0297	0.9726	0.0026	0.0301
15	19	0.0295	0.9540	0.0029	0.0297	0.9472	0.0028	0.0271
16	25	0.0295	0.9178	0.0034	0.0297	0.8995	0.0032	0.0232
17	21	0.0258	0.9122	0.0049	0.0241	0.9462	0.0037	0.0189
18	10	0.0123	1.2036	0.0049	0.0148	1.0679	0.0034	0.0150
19	09	0.0114	0.8627	0.0039	0.0093	1.0342	0.0027	0.0123
20	03	0.0000	–	–	0.0047	1.4784	0.0030	0.0113

Note: The bootstrap estimate $\hat{f}^*_k$ and $\bar{f}^*_k$ are generated by taking random samples of size 642 from the fitted probability densities, $\hat{f}$ and $\bar{f}$, respectively. Means and standard errors are calculated from 5000 repetitions. $\tilde{f}$ is the parametric estimate by fitting 3-term Fourier series to the grouped data, which is suggested in Buckland et al. (1992).

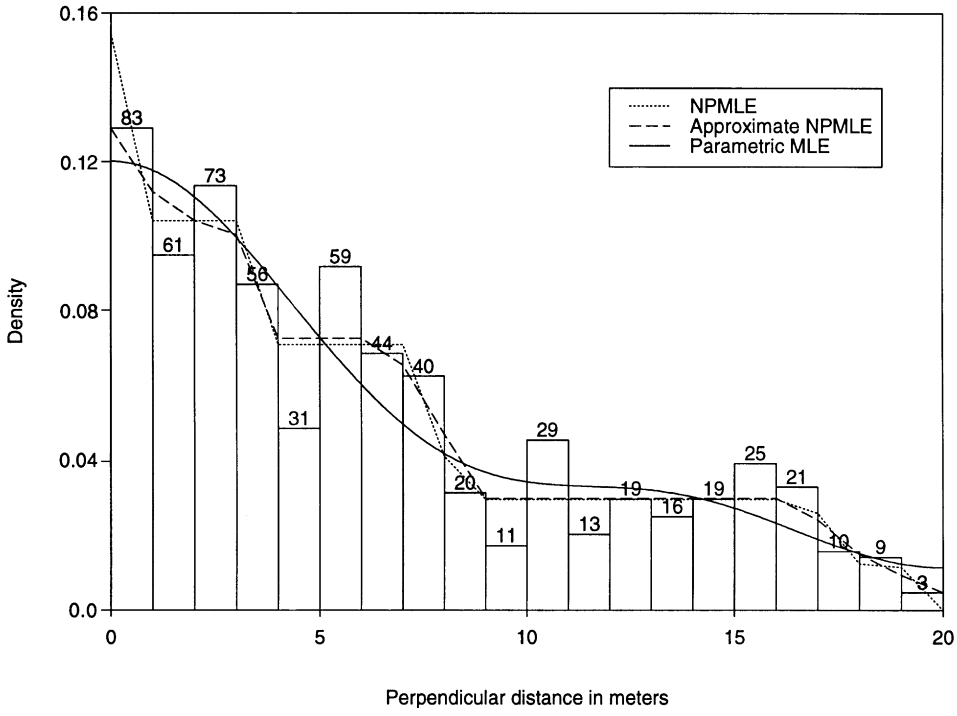

Fig. 1. Histogram of the wooden stake data. Also shown are the three estimated density functions.

obtained by fitting a 3-term Fourier series to the grouped data. This model is suggested by Buckland et al. (1992). Fig. 1 shows the comparison of the three density estimates.

Distance sampling is often used to investigate the population or population density in an area. In this case, 150 wooden stakes were set out in a 1000 m × 40 m rectangular area, and 11 samples were obtained by 11 observers. We work on the pooled data with size 642. A population density estimator is developed by Buckland et al. (1992), $\hat{D} = E(n) \times \hat{f}(0)/2L$, where L is the length of the field. The estimated population density is 45.03 stakes per hectare from $\hat{f}(0)$, 37.73 from $\bar{f}(0)$, and 35.13 from $\tilde{f}(0)$. Compared to the actual density, 37.5 stakes per hectare, $\hat{f}(0)$ overestimates $f(0)$. For ungrouped data, it is known that $\hat{f}(0)$ tends to be too large, and that problem appears to be present in this example also. A penalty term could be added to $f(0)$ easily following Woodroofe and Sun (1993).

3. Convergence

The proof that f^q converges to $\hat{f}$ is presented in this section. Let

$$\psi(f) = \sum_{k=1}^{m} n_k \log(f_{k-1} + f_k) - n \sum_{k=0}^{m} f_k \Delta y_k$$

for $f \in \Omega$. Then $\hat{f}$ maximizes ψ uniquely in Ω_0. In fact, $\hat{f}$ maximizes ψ uniquely in Ω. For, if $f \in \Omega$ and $\psi(f) > -\infty$, then $c = \sum_{k=0}^{m} f_k \Delta y_k > 0$, $g = f/c \in \Omega_0$, and $\psi(f) = \psi(g) + n[\log(c) - (c-1)] \leqslant \psi(g) \leqslant \psi(\hat{f})$ with strict inequality unless $f = \hat{f}$.

If $f \in \Omega$, $f_{m-1} > 0$, let $\psi_1(f;g) = \lim_{\varepsilon \downarrow 0}[\psi(f+\varepsilon g) - \psi(f)]/\varepsilon$ for $g \in \mathbb{R}^{m+1}$. Then

$$\psi_1(f;g) = \sum_{k=1}^{m} n_k \frac{g_{k-1}+g_k}{f_{k-1}+f_k} - n\sum_{k=0}^{m} g_k \Delta y_k.$$

Also, let $\Omega_f = \{g \in \mathbb{R}^{m+1}: f + \varepsilon g \in \Omega, \text{ for some } \varepsilon > 0\}$, and observe that if $g \in \Omega_f$, then $f + \varepsilon g \in \Omega$ for all sufficiently small $\varepsilon > 0$.

Lemma 1. *$\hat{f}$ is the unique $f \in \Omega$ for which $\psi_1(f;g) \leqslant 0$ for all $g \in \Omega_f$.*

Proof. It is clear that $\hat{f}_{m-1} > 0$ and $\hat{f}$ has the property claimed, since ψ is maximized at $f = \hat{f}$. To see that $\hat{f}$ is the only such function, suppose that $\tilde{f}$ has the property described in the Lemma. Let $g = \hat{f} - \tilde{f}$. Then $\tilde{f} + \varepsilon g = (1-\varepsilon)\tilde{f} + \varepsilon\hat{f} \in \Omega$ for all $0 \leqslant \varepsilon \leqslant 1$, $\psi(\tilde{f} + \varepsilon g)$ is concave in $0 \leqslant \varepsilon \leqslant 1$, and

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\psi(\tilde{f}+\varepsilon g) \leqslant \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\psi(\tilde{f}+\varepsilon g)|_{\varepsilon=0} = \psi_1(\tilde{f};g) \leqslant 0$$

for $0 \leqslant \varepsilon \leqslant 1$. So, the value of $\psi(\tilde{f} + \varepsilon g)$ at $\varepsilon = 1$ is at most the value at $\varepsilon = 0$. That is, $\psi(\hat{f}) \leqslant \psi(\tilde{f})$ and, therefore, $\tilde{f} = \hat{f}$. □

Now let $f \in \Omega_0$ be a fixed point of M; that is, $M(f) = f$. Further, let F be the continuous piecewise linear function for which $F(0) = 0$ and $F(y_k) = f_0\Delta y_0 + \cdots + f_k \Delta y_k$ for $k = 0, \ldots, m$. Then F is the least concave majorant of G_f, and it follows easily that

$$\sum_{j=0}^{m} f_j \Delta y_j = 1 \tag{9a}$$

and

$$\sum_{j=1}^{k} n_j + \left(\frac{f_k}{f_k + f_{k+1}}\right) n_{k+1} \leqslant n\sum_{j=0}^{k} f_j \Delta y_j \tag{9b}$$

for all $k = 0, \ldots, m-1$, with equality if $f_{k+1} < f_k$.

Lemma 2. *If $f \in \Omega_0$, $M(f) = f$, and either $f_m > 0$ or $f_{m-1} \geqslant n_m/n\Delta y_m$, then $f = \hat{f}$.*

Proof. It suffices to show that $\psi_1(f;h) \leqslant 0$ for all $h \in \Omega_f$. For $k = 0, \ldots, m$, let

$$g_j^k = f_j \quad \text{if } j \leqslant k,$$
$$g_j^k = 0 \quad \text{if } j > k.$$

Then $\psi_1(f;g^k) \leqslant 0$ for all k with equality if either $k = m$ or $k < m$ and $f_{k+1} < f_k$ by (9). The remainder of the argument is slightly different in the two cases $f_m > 0$ and $f_m = 0$.

Suppose first that $f_m > 0$. Then $g^0, \ldots, g^m$ form a basis for $\mathbb{R}^{m+1}$. So, any $h \in \mathbb{R}^{m+1}$ may be written as $h = \alpha_0 g^0 + \cdots + \alpha_m g^m$ in which case $h_k = \alpha_k f_k + \cdots + \alpha_m f_m$. If $\alpha_k < 0$ and $f_{k+1} = f_k$ for some $k < m$, then $f_k + \varepsilon h_k - (f_{k+1} + \varepsilon h_{k+1}) = \varepsilon\alpha_k f_k < 0$ and, therefore, $h \notin \Omega_f$. That is, if $h \in \Omega_f$, then $\alpha_k \geqslant 0$ whenever $f_{k+1} = f_k$. It follows that

$$\psi_1(f;h) = \sum_{j=0}^{m} \alpha_j \psi_1(f;g^j) \leqslant 0$$

and, therefore, that $f = \hat{f}$.

Suppose next that $f_m = 0$ and $f_{m-1} \geqslant n_m/n\Delta y_m$, and let $e = (0,\ldots,0,1)$. Then $\psi_1(f; g^{m-1}) = 0$, since $f_m < f_{m-1}$, and

$$\psi_1(f; e) = \frac{n_m}{f_{m-1}} - n\Delta y_m \leqslant 0.$$

In this case $g^0,\ldots,g^{m-1},e$ are a basis for $\mathbb{R}^{m+1}$. So, any $h \in \mathbb{R}^{m+1}$ may be written as $h=\alpha_0 g^0+\cdots+\alpha_{m-1}g^{m-1}+\beta e$. If $h \in \Omega_f$, then $\alpha_k \geqslant 0$ whenever $k < m-1$ and $f_{k+1} = f_k$, as above, and $\beta \geqslant 0$ (since $f_m = 0$). So, if $h \in \Omega_f$, then $\psi_1(f; h) = \sum_{k=0}^{m-1} \alpha_k \psi_1(f; g^k) + \beta\psi_1(f; e) \leqslant 0$, as above, and this implies that $f = \hat{f}$. □

Theorem. *If* $f_m^0 > 0$, *then* $\lim_{q\to\infty} f^q = \hat{f}$.

Proof. From Theorem 1 of Dempster et al. (1977), $l[M(f)] \geqslant l(f)$ with equality only if the conditional distributions of $r_{00},\ldots,r_{1m}$ are the same under $M(f)$ and f. In view of the nature of the conditional distributions and (4), this requires that $M(f) = f$. Thus, any f for which $l[M(f)] = l(f)$ is a fixed point of M. Clearly, $l(f^q)$ is non-decreasing in $q = 0, 1, \ldots$, and

$$L = \lim_{q\to\infty} l(f^q) \leqslant l(\hat{f})$$

exists and is finite. Since the parameter space Ω_0 is compact, it suffices to show that $\hat{f}$ is the only limit point of f^q, $q = 0, 1, 2, \ldots$, and for this is suffices to show that $L = l(\hat{f})$, since $l(f) = L$ for any limit point f, and l attains its maximum only at $\hat{f}$. Two cases are considered.

If $\limsup_{q\to\infty} f_m^q > 0$, then there is a limit point f of f_m^q, $q = 0, 1, 2, \ldots$ for which $f_m > 0$. It then follows from Lemma 2 that $f = \hat{f}$ and, therefore, that $L = l(\hat{f})$. Next, suppose that $\lim_{q\to\infty} f_m^q = 0$. In this case, it is easily seen that $f_m^q > 0$ for all q and $f_m^q < f_{m-1}^q$ for all sufficiently large q, since f_{m-1}^q cannot converge to zero. Thus,

$$f_m^{q+1} = \left(\frac{f_m^q}{f_{m-1}^q + f_m^q}\right)\frac{n_m}{n\Delta y_m}$$

for all sufficiently large q. Clearly, the limit inferior of f_m^{q+1}/f_m^q as $q \to \infty$, is at most one and, therefore,

$$\limsup_{q\to\infty} f_{m-1}^q \geqslant \frac{n_m}{n\Delta y_m}.$$

So, the sequence f^q, $q = 0, 1, 2, \ldots$ has a limit point f for which $f_{m-1} \geqslant n_m/(n\Delta y_m)$. As above, it then follows from Lemma 2 that $f = \hat{f}$ and, therefore, that $L = l(\hat{f})$. □

References

Bickel, P.J., Fan, J., 1996. Some problems on the estimation of unimodal densities. Statist. Sinica 6, 23–45.

Buckland, S., Anderson, D., Burnham, K., Laake, J., 1992. Distance Sampling. Chapman & Hall, London.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Statist. Soc. B 39, 1–22.

Johnson, E., Routledge, R., 1985. The line transect method: a nonparametric estimator based on shape restrictions. Biometrics 41, 669–679.

Robertson, T., Wright, F., Dykstra, R., 1989. Order Restricted Statistical Inference. Wiley, New York.

Woodroofe, M., Sun, J., 1993. A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. Statist. Sinica 3, 501–515.