Statistica Sinica **7**(1997), 545-575

SEMI-PARAMETRIC ESTIMATES UNDER BIASED SAMPLING

Jiayang Sun and Michael Woodroofe

University of Michigan

Abstract: In observational studies subjects may self select, thereby creating a biased sample. Such problems arise frequently, for example, in astronomical, biomedical, animal, and oil studies, survey sampling and econometrics. For a typical subject, let Y denote the value of interest and suppose that Y has an unknown density function f. Further, let $w(y)$ denote the probability that the subject includes itself in the study given $Y = y$. Then the conditional density of Y given that it is observed is $f^*(y) = w(y)f(y)/\kappa$, where κ is a normalizing constant. The problem of estimating w and f from a biased sample $X_1, \ldots, X_n$ independently from f^* is considered when f is known to belong to a parametric family, say $f = f_\theta$, where θ is a vector of unknown parameters, and w is assumed to be non-decreasing. An algorithm for computing the maximum likelihood estimator of (w, θ) is developed, and consistency is established. Simulations are used to show that our method is feasible with moderate sample size, and applications to animal and oil data are given.

Key words and phrases: Animal and oil data, convergence, consistency, EM algorithm, incomplete sample, maximum likelihood estimates, MM algorithm, selection bias, simulations.

1. Introduction

The Problem. Consider a multiparameter exponential family of univariate densities with respect to a measure Λ, say

$$f_\theta(y) = \exp\{\theta' T(y) - \psi(\theta)\}, \qquad \text{for all } y \in \mathcal{Y}, \quad \theta \in \Omega, \tag{1}$$

where $\Omega \subseteq \mathbb{R}^p$, $\mathcal{Y} \subseteq \mathbb{R}$ and $T : \mathcal{Y} \to \mathbb{R}^p$ for some $p \geq 1$. Suppose that $Y_1, \ldots, Y_N$ are i.i.d. with common density f_θ for some unknown $\theta \in \Omega$ and that Y_i is observed with probability $w(Y_i)$ given Y_i, where w is an unknown function on $\mathcal{Y}$. If $X_1, \ldots, X_n$ denote the observed values, then $X_1, \ldots, X_n$ may be regarded as a biased sample from f_θ. The problem considered here is to estimate (w, θ) from such a biased sample. Problems of this nature may arise in several ways.

Examples

1. *Complaint Data.* If a manufacturer distributes N lots of its product to customers, then it may be reasonable to suppose that the number of defectives,

Y_i say, in the ith lot has a Poisson distribution and that the probability that the user reports the number of defectives to the manufacturer is a non-decreasing function of Y_i.

2. *Sample surveys.* If a questionnaire is mailed to each of N randomly selected individuals who are asked to supply their own values of a variable, Y say, then the probability that a given individual responds may well depend on Y.

3. *Astronomy.* In the discussion of Lynden-Bell (1992), Woodroofe (1992) describes an example in which Y is the (observable) angular diameter of a galaxy (in suitable units), and there is an unknown, non-decreasing selection function w.

4. *Animal Studies.* Patil and Rao (1977) discussed an example in which Y is the group size of moose in northeast Minnesota. Due to the visibility bias, groups with smaller size have smaller probability to be seen.

5. *Oil.* If Y is the volume (size) of an oil field, then it is easier to find larger Y's than small ones. See, for example, Gordon (1993) and Bloomfield et al. (1979).

There are similar examples in reliability and biomedical research and econometrics. In biomedical research censoring may occur besides the bias. As the examples indicate, the value of N may be known (Examples 1 and 2) or unknown (Examples 3-5). The estimation of w and θ differs in the two cases.

Of course, some restriction must be placed on w in order to insure identifiability. In previous (non-regression) papers, w is often assumed known or to be a known parametric function with some unknown parameters. See, for example, Vardi (1982, 1985), Gill, Vardi and Wellner (1988), Robbins and Zhang (1988), Gordon (1993), Bloomfield et al. (1979), and Vardi and Zhang (1992) et al. Here the case in which w is a non-decreasing function is considered. This condition seems reasonable in Examples 1, and 3-5, and may be reasonable in some cases in Example 2. Our model is complementary to the earlier work, and may be useful in assessing the validity of parametric assumptions on w. We shall use the animal and oil data to illustrate these points. There is some related work in econometrics, where w is assumed to be smooth and estimated by kernel methods. See Manski (1993) and the references therein. These methods differ from ours since they postulate some possibly incomplete data from the unobserved subjects.

In Section 3, the maximum likelihood estimators and penalized maximum likelihood estimators of (w, θ) are developed through an MM (maximization-maximization) algorithm, when N is known or unknown. The MM algorithm is similar to the EM algorithm in concept but differs in an important way: the maximizing value may be a boundary point (cf. Wu (1983)). In Section 4, applications to oil and animal data are given as examples, and simulations for various models are performed to examine our methods in the finite sample situation. A general convergence theorem for the MM algorithm and its application to our

(penalized) maximum likelihood estimates are given in Section 5. The consistency of the estimators is shown in Section 6. Amusingly, convergence of the MM algorithm is more difficult for known N, and consistency is more difficult for unknown N. A penalty term with a proper smoothing parameter has to be introduced to insure the consistency in the case of unknown N. The proofs differ from those of Vardi and Zhang (1992) for empirical distributions. Some concluding remarks about the asymptotic distribution and choice of the penalizing parameter are given in Section 7. An optimization theorem and its proof are provided in the Appendix. These may be of independent interests. Some preliminaries are presented in Section 2.

2. Preliminaries

The Model. To formalize the problem, let W denote the set of all nondecreasing functions $w : \mathcal{Y} \to (0, 1]$ for which $\int_{\mathcal{Y}} w d\Lambda > 0$. If Y has density f_θ, where $\theta \in \Omega$, and Y is observed with probability $w(Y)$, where $w \in W$, then the probability that Y is observed and its conditional density given the observations are respectively

$$\kappa(w, \theta) = \int_{\mathcal{Y}} w(y) f_\theta(y) \Lambda(dy) \quad \text{and} \quad f^*_{w,\theta}(y) = \frac{w(y) f_\theta(y)}{\kappa(w, \theta)}$$

for all $y \in \mathcal{Y}$. (Here $\kappa(w, \theta) > 0$ for all $(w, \theta) \in W \times \Omega$.) Then the biased sample introduced in Section 1 may be modeled by random variables n and $X_1, \ldots, X_n$, where $n \sim \text{Binomial}[N, \kappa(w, \theta)]$ and $X_1, \ldots, X_n$ are drawn independently from $f^*_{w,\theta}$, given n. Here $(w, \theta) \in W \times \Omega$ is unknown and N may be known or unknown. In addition, there may be an independent (of $n, X_1, \ldots, X_n$) sample from f_θ, say $Y_1, \ldots, Y_m$, where $m \geq 0$. The complete sample is absent if $m = 0$. This model differs from those of Bickel, Nair and Wang (1992), and Gordon (1993) where the population is assumed to be finite.

Some Conditions. In (1), Ω is taken to be the natural parameter space of the family and is assumed to be open. In addition, two special properties are required of the family. Let $\mathcal{Y}$ denote the support of Λ. It is required that $\mathcal{Y}$ be unbounded on the right; that is, $\sup \mathcal{Y} = \infty$. This requires $\mathcal{Y}$ to contain infinitely many points, but little else in view of the generality of T. Next, it is required that the family of conditional distributions given $Y \geq x$ be minimal for each $x \in \mathcal{Y}$, using the terminology of Brown (1986). Letting Λ^x be the restriction of Λ to $[x, \infty)$, the condition may be written

$$\text{dimension } \{\text{convexhull } [\text{support}(\Lambda^x \circ T^{-1})]\} = p, \qquad \text{for all } x \in \mathcal{Y}. \tag{2}$$

With $p = 1$, the conditions are satisfied by the exponential, geometric, normal, and Poisson families, but not by the binomial. They are also satisfied by the two parameter normal and gamma families.

Properties. For any fixed $w \in W$, it is easy to see that $\{f^*_{w,\theta} : \theta \in \Omega\}$ is another exponential family of the form (1), with Λ replaced by $\Lambda_w(dy) = w(y)\Lambda(dy)$ and ψ by

$$\psi_w(\theta) = \log\Big\{\int_{\mathcal{Y}} \exp[\theta' T] d\Lambda_w\Big\} \leq \infty, \qquad \theta \in \mathbb{R}^p,$$

and this family is minimal by (2). The natural parameter space for the latter family, $\Omega_w = \{\theta \in \mathbb{R}^p : \psi_w(\theta) < \infty\}$ contains Ω, and the inclusion may be proper. With this notation,

$$\kappa(w, \theta) = \exp[\psi_w(\theta) - \psi(\theta)] \tag{3}$$

for all $\theta \in \Omega$, $w \in W$.

For any $\theta \in \Omega$, the mean $\nabla\psi(\theta)$ of f_θ is in the expectation space $\nabla\psi(\Omega)$, obviously, where ∇ denotes gradient with respect to θ. The next lemma shows that the mean $\nabla\psi_w(\theta)$ of $f^*_{w,\theta}$ is in $\nabla\psi(\Omega)$ too for any $w \in W$.

Lemma 1. *If $w \in W$, then $\nabla\psi_w(\Omega^\circ_w) \subseteq \nabla\psi(\Omega)$, where Ω°_w denotes the interior of Ω_w.*

Proof. From Brown (1986), it is known that$\nabla\psi_w(\Omega^\circ_w)$ is contained in the interior of the convex support of $\Lambda_w \circ T^{-1}$ for each $w \in W$ with equality when $w = 1$, in which case the family is steep. Thus, it suffices to show that support$(\Lambda_w \circ T^{-1}) \subseteq$ support$(\Lambda \circ T^{-1})$ for all $w \in W$; and this is clear, since Λ_w is absolutely continuous with respect to Λ.

Lemma 2. *Let J be any compact subset of $\nabla\psi(\Omega)$. If $\theta^j \in \Omega$, $j \geq 1$, and either $||\theta^j|| \to \infty$ or $\theta^j \to \theta \notin \Omega$ as $j \to \infty$, then*

$$\limsup_{j\to\infty} \sup_{t\in J} \; [t'\theta^j - \psi(\theta^j)] = -\infty.$$

Proof. If $\theta^j \to \theta \notin \Omega$, then the result is clear, since then $\psi(\theta) = \infty$ and $\liminf_{j\to\infty} \psi(\theta^j) \geq \psi(\theta)$, by Fatou's Lemma. If $||\theta^j|| \to \infty$ as $j \to \infty$, then the result follows easily from Lemma 3.5 and Lemma 5.3 of Brown (1986), pp. 73-74, pp. 146.

More General Models. The model described above may be used as a building block in more complicated ones. For example, in many case control studies, there is a $\{0, 1\}$ valued covariate Z and the conditional distribution of an observable X given $Z = j$ is of the form: $w_j(x) f_{\theta_j}(x)/[\kappa(w_j, \theta_j)]$, for $j = 0, 1$. If w_0 and w_1 are assumed to be equal (monotone or not), then

$$P_{\theta_0,\theta_1}[Z = 1|X = x] = \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)},$$

where $\beta = \theta_1 - \theta_0$ and $t = T(x)$, and inference about β may be carried out by logistic regression. (See, for example, Breslow (1980).) The methods advocated in this paper may be useful for estimating the individual θ_j and, in principle, for checking the assumption that $w_0 = w_1$.

3. Maximum Likelihood Estimators

In the derivation of the maximum likelihood estimators, suppose that $1 \leq n < N$ and denote the observed order statistics of $X_1, \ldots, X_n$ by $x_1 < \cdots < x_r$. Further, let $n_i = \#\{k \leq n : X_k = x_i\}$ for $i = 1, \ldots, r$, and

$$\bar{t} = \frac{\sum_{i=1}^{n} T(X_i) + \sum_{j=1}^{m} T(Y_j)}{m+n}.$$

It is assumed that $\bar{t}$ is an interior point of the expectation space of the family; that is, $\bar{t} \in \nabla\psi(\Omega)$. By Lemma 1, this holds with probability approaching one as $N \to \infty$.

Known N. If N is known, then the log likelihood function is

$$\begin{aligned} l(w,\theta) &= \sum_{i=1}^{r} n_i \log[w(x_i)] + (m+n)[\theta'\bar{t} - \psi(\theta)] + (N-n)\log[1-\kappa(w,\theta)] \\ &= \sum_{i=1}^{r} n_i \log[w(x_i)] + (m+n)\theta'\bar{t} + (N-n)\log[e^{\psi(\theta)} - e^{\psi_w(\theta)}] - (N+m)\psi(\theta) \quad (4) \end{aligned}$$

for $\theta \in \Omega$ and $w \in W$, using (3). It is easily seen that for any fixed $\theta \in \Omega$, $l(w,\theta)$ is maximized when w is a step function of the form $w(y) = 0$ for $y < x_1$ and $w(y) = w_k$ for $x_k \leq y < x_{k+1}, k = 1, \ldots, r$, where $0 \leq w_1 \leq \cdots \leq w_r \leq 1$ and $x_{r+1} = \infty$. Then $w(x_k) = w_k, k = 1, \ldots, r$, in (4), and

$$\kappa(w,\theta) = \sum_{k=1}^{r} p_k(\theta) w_k, \qquad \text{for all } \theta \in \Omega,$$

where

$$p_k(\theta) = \int_{[x_k, x_{k+1})} f_\theta(y)\Lambda(dy), \quad k = 0, \ldots, r,$$

with $x_0 = -\infty$. So the maximization may be restricted to such step functions.

It is straightforward to maximize the likelihood function with respect to one variable when the other is held fixed. For fixed θ, maximizing $l(w,\theta)$ with respect to w is an exercise in isotonic estimation (cf. Robertson, Wright and Dykstra (1988)), and the maximizing values are

$$\hat{w}_k(\theta) = \min[\frac{n(1-c)}{N-n}\tilde{w}_k(\theta), \ \ 1], \quad k = 1, \ldots, r, \qquad (5)$$

where

$$\tilde{w}_k(\theta) = \max_{i \le k} \min_{k \le j \le r} \frac{n_i + \cdots + n_j}{n(p_i(\theta) + \cdots + p_j(\theta))}, \quad k = 1, \ldots, r,$$

and $0 < c = c(\theta) \le 1$ is the unique solution to the equation

$$c = \sum_{k=1}^{r} p_k(\theta) \min[\frac{n(1-c)}{N-n}\tilde{w}(\theta), \ \ 1].$$

(See the Appendix for the derivation.) Here $\kappa[\hat{w}(\theta), \theta] = c$, and $c \le n(1-c)/(N-n)$ from (A.7) in the Appendix. So, $c \le n/N$ and, therefore,

$$\kappa[\hat{w}(\theta), \theta] \le \frac{n}{N} \tag{6}$$

for all $\theta \in \Omega$. This relation is needed in the proof of convergence in Section 5.

Conversely, for fixed w, $l(w, \theta)$ attains its maximum with respect to θ, at a point $\hat{\theta}(w)$ for which $\nabla l(w, \theta) = 0$, since $\theta'\bar{t} - \psi(\theta) \to -\infty$ as either $||\theta|| \to \infty$ or θ approaches a boundary point of Ω. Differentiating the second line of (4), the latter equation may be written

$$\frac{N+m}{n+m}\nabla\psi(\hat{\theta}(w)) + \frac{N-n}{n+m}\left[\frac{\kappa(w, \hat{\theta}(w))\nabla\psi_w(\hat{\theta}(w)) - \nabla\psi(\hat{\theta}(w))}{1 - \kappa(w, \hat{\theta}(w))}\right] = \bar{t}. \tag{7}$$

As shown in Section 5, the true maximum likelihood estimators of (w, θ) are in the set $\{(w, \theta) \in W \times \Omega : \kappa(w, \theta) \le (n+m)/(N+m)\}$ and l is concave on this set. So, any solution of (7) in the set is a maximum point of l with respect to θ given w.

These two special cases suggest an iterative procedure. Let $\hat{w}^0$ denote an initial guess–for example, $\hat{w}^0 = 1$; and let

$$\hat{\theta}^k = \hat{\theta}(\hat{w}^{k-1}) \qquad \text{and} \qquad \hat{w}^k = \hat{w}(\theta^k) \tag{8}$$

for $k = 1, 2, \ldots$, where $\hat{w}(\theta)$ and $\hat{\theta}(w)$ are defined by (5) and (7). It is shown in Section 5 that the sequence is precompact and that any limit point is a maximum likelihood estimator.

Unknown N. If N is unknown, then the conditional log likelihood function given n is

$$\begin{aligned} l^*(w, \theta) &= \sum_{i=1}^{r} n_i \log[w(x_i)] + (m+n)[\theta'\bar{t} - \psi(\theta)] - n\log[\kappa(w, \theta)] \\ &= \sum_{i=1}^{r} n_i \log[w(x_i)] + (m+n)\theta'\bar{t} - [m\psi(\theta) + n\psi_w(\theta)] \end{aligned} \tag{9}$$

for $\theta \in \Omega$ and $w \in W$, using (3) again. It is clear that $l^*(cw, \theta) = l^*(w, \theta)$ for all $w \in W$, $\theta \in \Omega$ and $c > 0$ for which $cw \leq 1$. Thus, there cannot be a unique maximum likelihood estimate for w. The lack of uniqueness foreshadows problems with consistency. To overcome these, a penalized log-likelihood of the form

$$l^*_\alpha(w, \theta) = l^*(w, \theta) - \frac{\alpha n}{\kappa(w, \theta)} \tag{10}$$

is considered ($l^*_\alpha = l^*$ if $\alpha = 0$), where $0 < \alpha = \alpha_n \leq 1$ may approach zero as $n \to \infty$, as in Woodroofe and Sun (1993). In (10), the log-likelihood function has been penalized (made smaller) for small values of $\kappa(w, \theta)$, the probability of observing an X. This term has been included to force some regularity of the estimators. To the best of our knowledge, the use of the term "penalized" to describe this process originated with the work of Good and Gaskins (1971). The penalized log-likelihood is maximized subject to the constraints

$$w(y) \geq \epsilon, \quad \text{for all } y \in \mathcal{Y} \qquad \text{and} \qquad \sup_y w(y) = 1,$$

where $\epsilon > \alpha$ represents a lower bound for the probability of observing a Y, given its value. The condition that $\sup_y w(y) = 1$ represents no real restriction.

Note that $\kappa(w, \theta) \geq \epsilon$ since $w \geq \epsilon$ and that $-\log \kappa - \alpha/\kappa$ decreases in κ when $\kappa > \alpha$. So, as above, for any $\theta \in \Omega$, $l^*_\alpha(w, \theta)$ is maximized when w is a step function of the form $w(y) = \epsilon$ for $y < x_1$ and $w(y) = w_k$ for $x_k \leq y < x_{k+1}, k = 1, \ldots, r$, where $\epsilon \leq w_1 \leq \cdots \leq w_{r-1} \leq w_r = 1$. Then $w(x_k) = w_k, k = 1, \ldots, r$, in (10), and

$$\kappa(w, \theta) = \epsilon p_0(\theta) + \sum_{k=1}^{r-1} p_k(\theta) w_k + p_r(\theta)$$

for all $\theta \in \Omega$.

It is again easy to maximize the (penalized) likelihood function with respect to either variable when the other is held fixed. For fixed θ, the maximizing values are $\hat{w}_r(\theta) = 1$ and

$$\hat{w}_k(\theta) = \max\left\{\epsilon, \min[\frac{c^2}{c - \alpha}\tilde{w}_k(\theta), 1]\right\}, \quad k = 1, \ldots, r-1, \tag{11}$$

where

$$\tilde{w}_k(\theta) = \max_{i \leq k} \min_{k \leq j < r} \frac{n_i + \cdots + n_j}{n(p_i(\theta) + \cdots + p_j(\theta))}, \quad k = 1, \ldots, r-1,$$

and $0 < c = c(\theta) < 1$ is the largest solution to the equation

$$c = \epsilon p_0(\theta) + \sum_{k=1}^{r-1} p_k(\theta) \max\{\epsilon, \ \min[\frac{c^2}{c - \alpha}\tilde{w}_k, \ 1]\} + p_r(\theta). \tag{12}$$

See the Appendix for details.

Now consider a fixed w for which $w_1 \geq \epsilon > \alpha \geq 0$ and, therefore, $w_k \geq \epsilon$ for all $k = 1, \ldots, r$. Then l_α^* is strictly concave with respect to $\theta \in \Omega$. This may be checked by differentiation using $\kappa(w, \theta) \geq \epsilon > \alpha$. (See (18) in Section 5 for a more general result.) It follows that $l_\alpha^*(w, \theta)$ attains its maximum at a unique value, $\hat{\theta}(w)$ say, and that $\hat{\theta}(w)$ is the unique solution to the equation $\nabla l_\alpha^*[w, \theta(w)] = 0$. Differentiating the second line of (9), the latter equation may be written

$$\frac{m}{m+n}\nabla\psi[\hat{\theta}(w)] + \frac{n}{m+n}\nabla\psi_w[\hat{\theta}(w)] - \frac{\alpha n}{m+n}\left[\frac{\nabla\psi_w(\hat{\theta}(w)) - \nabla\psi(\hat{\theta}(w))}{\kappa(w, \hat{\theta}(w))}\right] = \bar{t}. \tag{13}$$

The iterative algorithm (8) is again suggested, and it may be shown that $(\hat{w}_k, \hat{\theta}_k)$ converges to the penalized maximum likelihood estimator. (See Section 5.)

4. Examples and Simulations

In this section, we present two examples and some simulations. The first example is for a Poisson model which is discrete with a one-dimensional parameter, and the second is for a (Log)normal model which is continuous with a two-dimensional parameter. Two precisions are to be specified in applying our procedure. One is for the (penalized) Maximum likelihood estimate of θ given w and that for w given θ, the inner loop. The other is for the MM algorithm, the outer loop. The inner precision should not be set too large, since it may cause the outer loop to oscillate between two values.

Example 4 – *Revisited.* Aerial Moose Census Data. The moose data from northeast Minnesota listed in Table 1 were collected by James Peek, University of Minnesota, in 1969. Columns 1 and 2 give the group size and counts of moose, with total 113 groups. Let $Y = \text{size}-1$ and $w(y)$ be the probability that a group of moose of size $y+1$ is seen. Following Cook and Martin (1974) and Patil and Rao (1977), we suppose that Y has a Poisson distribution with an unknown intensity parameter $\theta > 0$. We also assume that $w(y)$ is a non-decreasing function since a smaller group has a smaller probability to be seen. In fact, a parametric form of w is assumed in these two references. Specifically, in Cook and Martin (1974), $w(y) = 1 - q^{y+1}$ for some $0 < q < 1$, while in Patil and Rao, $w(y) = c(y+1)$ for some constant $c > 0$ (called *length biased* sampling). The other columns in Table 1 are corresponding estimates of Cook and Martin (C-M), Patil and Rao (P-R) with $c = 1/6$ and our estimates with $\alpha n = 0.4, 0.5, 0.6, 0.8$ and 2.0. The two precisions are 0.001 for the Moose data and our estimates required at most 12 iterations in the MM algorithm. For all choices of α, $\hat{w}(1) < \hat{w}(2)$, but $\hat{w}(k)$ are nearly constant for $k = 2, \ldots, 5$. This is consistent with Cook and Martin, but not with Patio and Rao (cf. Table 1 and Figure 1).

Table 1. The group size of moose

		$\hat{w}$						
size	counts	$\alpha n = 0.4$	$\alpha n = 0.5$	$\alpha n = 0.6$	$\alpha n = 0.8$	$\alpha n = 2$	C-M	P-R
1	45	0.4182	0.5268	0.6270	0.8225	0.9287	0.89	0.1667
2	46	0.4933	0.6092	0.7183	0.9327	1.0000	0.987	0.3333
3	15	0.4933	0.6092	0.7183	0.9327	1.0000	0.99867	0.5000
4	5	0.5111	0.6131	0.7183	0.9327	1.0000	0.99985	0.6667
5	1	0.5111	0.6131	0.7183	0.9327	1.0000	0.99998	0.8333
6	1	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000
	$\hat{\theta}$	0.8152	0.8280	0.8336	0.8397	0.8587	0.84	0.505
	$\hat{\kappa}$	0.4617	0.5741	0.6791	0.8852	0.9698		

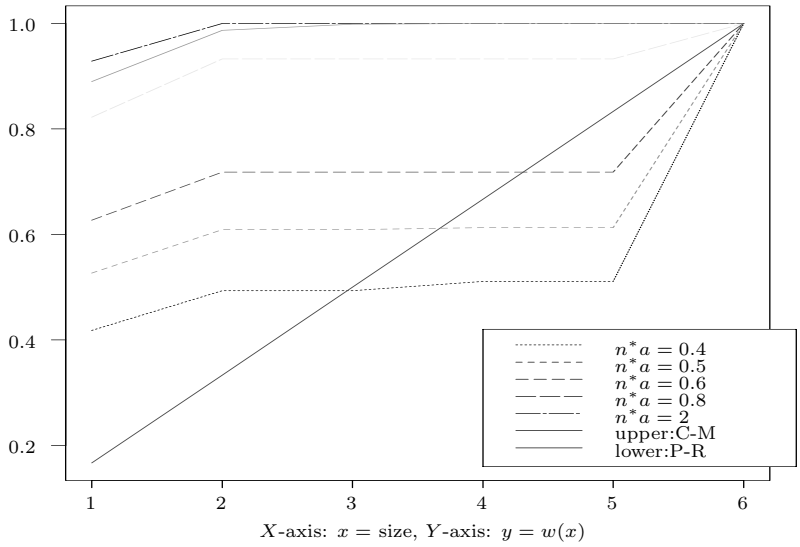

Figure 1. Estimates of $w(x)$ for moose data
Notes. These are unsmoothed plots, i.e. interpolating between two points.

Example 5– *Revisited*. Oil Data. The field size Y of 58 oil discoveries (in units of 10^6 BBLS) listed in Table 2 are from Meisner and Demirmen (1981). As indicated by both Bloomfield et al. (1979) and Meisner and Demirmen (1981), the study of $w(y)$, the probability of discovering a field of size y, is an important component in a model of forecasting future discoveries. In this case, it is reasonable to assume that Y has a Lognormal distribution, as in these two references. Bloomfield et al. (1979) supposed that $w(y)$ was a lower power function of y, rather than proportional to y, and reported a much better fit than the linear fit for the Kansas data therein. For the data of Meisner and Demirmen (1981), the sample mean and the estimated standard deviation of log(size) are respectively

3.848 and 1.332, using $s(\bar{t}, t)$ defined in (14) below. Several typical $\hat{w}$ are presented in Table 2. They required at most 13 iterations in the MM algorithm. The precisions for the inner and outer loops are 0.0001 and 0.001 throughout. It is clear from our $\hat{w}$ and Figure 2 that no power function of y fits the data well. Indeed, to compensate this, Meisner and Demirmen (1981) assume that $w(y)$ is a power function of y, but the power decreases as the number of the drillings increases.

Table 2. The size of oil fields

		$n\alpha =$	0.01	1	2				0.01	1	2
no	size	log(size)	$\hat{w}$			no	size	log(size)	$\hat{w}$		
1	5.9	1.775	0.238	0.439	0.596	...	...	...	...	...	...
...	...	...	...	...	...	53	337	5.820	0.984	1.000	
38	75	4.317	0.341	0.661	0.908	...	...	...	...	...	...
...	...	...	...	...	...	57	775	6.653	1.000	1.000	
45	125	4.828	0.478	0.819		...	...	...	...	...	...
...	...	...	...	...	...	58	1328	7.191	1.000	1.000	
46	154	5.037	0.497	0.825		$\hat{\mu}$		3.848	3.319	3.357	3.406
...	...	...	...	...	...	$\hat{\sigma}$		1.332	1.331	1.455	1.531
51	215	5.371	0.548	0.847	1.000	$\hat{\kappa}$			0.429	0.695	0.878

Notes. The dots indicate that the estimates are same as the next one. The $\hat{\mu}$ and $\hat{\sigma}$ are the estimates of the parameters in the Lognormal distribution of Y.

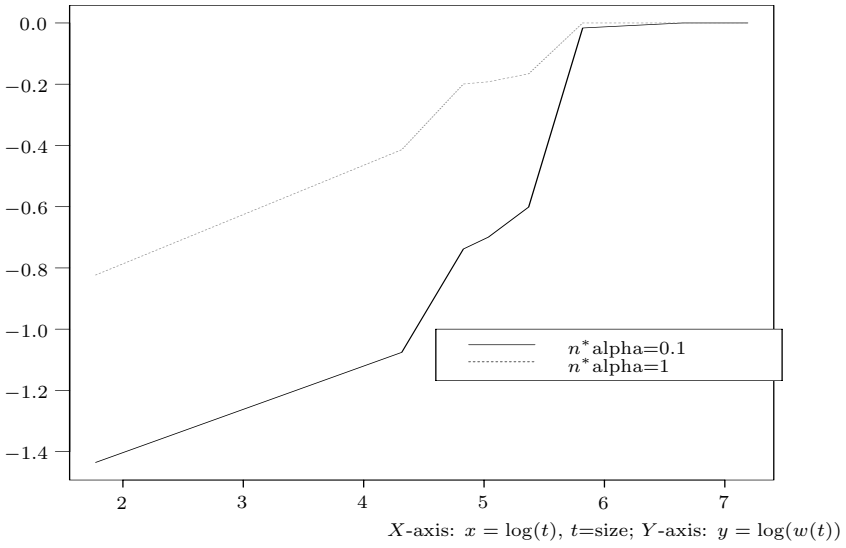

Figure 2. Estimates $\log(\hat{w}(x))$ for oil data

Notes. These are unsmoothed plots at log-scales. A straight line would indicate that some power function of y fitted w well.

Simulations. Simulation experiments were conducted for Normal and Poisson data, $n = 20, 50$ and 100, and $m = 0$ and $n/2$ for both known and unknown N. There were 1000 replications for each combination. Not all the results are reported in detail. The choice of the smoothing parameter α_n presents a difficult problem for the case of unknown N. Some possible approaches to this question are described in Section 7. In the simulation, α_n is chosen to reduce the bias of $\hat{\kappa}$. The desirable α_n for the case of $m = 0$ is bigger than that for $m = n/2$.

Poisson. Consider the case that Y has a Poisson distribution with mean μ, i.e. $Y \sim \text{Poisson}(\mu)$, and $w(y) = (y+1)/(y+4), y = 1, 2, \ldots$ Then the natural parameter is $\theta = \log(\mu)$. The true κ value, κ^o, say, can be computed by

$$\kappa^o = \sum_{i=0}^{\infty} \frac{i+1}{i+4} \frac{\mu^i}{i!} e^{-\mu},$$

which gives 0.4727 for $\mu = 2$ and 0.7738 for $\mu = 10$. When N is unknown the penalized maximum likelihood estimator $\hat{\theta}(w) = \log(\hat{\mu}(w))$ of θ given w is the solution of $e(\mu) = 0$ where

$$e(\mu) := \bar{t} - \mu - \mu \left[\frac{n}{m+n} - \frac{n\alpha}{m+n} \frac{1}{\kappa(w,\theta)} \right] \frac{\kappa'_\mu(w,\theta)}{\kappa(w,\theta)}$$

and κ'_μ is the partial derivative of κ with respect to μ. It is easy to see that $e(0) > 0$ and $e(\bar{t}) < 0$. Thus, a simple bisection algorithm can be used to solve this problem. When N is known the maximum likelihood estimator $\hat{\theta}(w)$ is a solution of $e_1(\mu) = 0$ where

$$e_1(\mu) := \bar{t} - \mu - \mu \left(\frac{N-n}{m+n} \right) \left[\frac{\kappa'_\mu(w,\theta)}{1 - \kappa(w,\theta)} \right].$$

Given μ or θ, calculating $\hat{w}$ follows from (5) or (11) directly, based on an inner loop precision. The simulation results are presented in Table 3. The same random numbers are used for two cases, known and unknown N, and precisions for the inner and outer loops are 0.001. With unknown N, the average iteration number for the MM algorithm is $7.3, 8.5, 9.4$ for $n = 20, 50, 100$, respectively, when $\mu = 2$; and $5.3, 6.7, 7.4$ when $\mu = 10$. The average iteration number is smaller when N is treated known. The results are fairly good even for $n = 20$. The estimate $\hat{\mu}$ is closer to the true μ when N is known. For bigger αn (than those used in the table), say $0.8, 1.2, 1.6$ for $n = 20, 50, 100$, respectively, and data from Poisson(2), the corresponding sample means of $\hat{\mu}$ are $2.001, 2.010, 2.022$ but those for $\hat{\kappa}$ are $0.72, 0.73, 0.74$ with smaller standard deviation than those reported in Table 3. Hence, a bigger α (than those in Table 3) increases the bias and decreases the variance of $\hat{\kappa}$.

Table 3. The summary statistics under the Poisson model

n	Poisson(2), $\kappa^o = 0.4727$						Poisson(10), $\kappa^o = 0.7738$					
	$m = n/2$			$m = 0$			$m = n/2$			$m = 0$		
	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
20	$n\alpha_n = 0.08$			$\alpha_n = 0.03n^{-0.5}$			$n\alpha_n = 1.2$			$\alpha_n = 0.24n^{-0.5}$		
N	42.00	42.41	6.876	42.00	42.39	6.711	26.00	25.90	2.689	26.00	25.96	2.771
$\bar{t}$	2.233	2.231	0.255	2.350	2.344	0.314	10.13	10.15	0.586	10.20	10.21	0.700
$\hat{\mu}$	1.883	1.882	0.288	1.684	1.706	0.338	9.734	9.718	0.636	9.357	9.320	0.765
	1.924	1.918	0.276	1.89	1.90	0.327	9.749	9.745	0.606	9.58	9.56	0.727
d	0.375	0.398	0.088	0.375	0.377	0.063	0.525	0.512	0.068	0.525	0.514	0.066
	0.300	0.317	0.060	0.300	0.308	0.062	0.625	0.608	0.074	0.625	0.607	0.076
$\hat{\kappa}$	0.432	0.470	0.232	0.337	0.375	0.166	0.782	0.772	0.142	0.674	0.680	0.123
	0.471	0.476	0.075	0.466	0.472	0.077	0.761	0.762	0.078	0.755	0.757	0.081
50	$n\alpha_n = 0.112$						$n\alpha_n = 1.34$					
N	105.0	105.3	10.49	106.0	106.0	10.64	64.00	64.50	4.178	64.00	64.65	4.237
$\bar{t}$	2.220	2.227	0.162	2.320	2.336	0.200	10.12	10.13	0.365	10.18	10.20	0.425
$\hat{\mu}$	1.916	1.915	0.201	1.854	1.863	0.294	9.738	9.751	0.415	9.489	9.502	0.513
	1.940	1.944	0.191	2.03	2.03	0.230	9.746	9.776	0.383	9.64	9.65	0.467
d	0.363	0.375	0.085	0.340	0.378	0.094	0.471	0.465	0.066	0.471	0.463	0.067
	0.288	0.295	0.045	0.250	0.269	0.032	0.571	0.557	0.073	0.571	0.555	0.076
$\hat{\kappa}$	0.438	0.474	0.224	0.375	0.463	0.235	0.782	0.774	0.138	0.733	0.737	0.117
	0.474	0.476	0.047	0.469	0.471	0.048	0.770	0.771	0.050	0.764	0.765	0.051
100	$n\alpha_n = 0.185$						$n\alpha_n = 1.63$					
N	212.0	212.1	15.22	211.0	211.8	15.09	129.0	129.3	6.059	129.0	129.3	6.033
$\bar{t}$	2.227	2.226	0.115	2.340	2.338	0.137	10.13	10.13	0.249	10.19	10.20	0.309
$\hat{\mu}$	1.943	1.945	0.160	1.991	1.977	0.261	9.798	9.793	0.307	9.627	9.618	0.400
	1.965	1.963	0.152	2.11	2.11	0.167	9.813	9.815	0.277	9.73	9.73	0.352
d	0.333	0.352	0.074	0.359	0.399	0.119	0.416	0.426	0.067	0.423	0.427	0.068
	0.273	0.282	0.037	0.250	0.256	0.015	0.500	0.516	0.077	0.500	0.518	0.077
$\hat{\kappa}$	0.449	0.472	0.206	0.467	0.565	0.274	0.785	0.774	0.133	0.778	0.774	0.107
	0.470	0.472	0.034	0.470	0.472	0.034	0.771	0.771	0.036	0.766	0.768	0.036

Notes. That d is the Kolmogorov-Sminov distance between $\hat{w}$ and w^o (the true value of w). The upper figures are the estimates when N is treated unknown and the lower are these when N is known.

Normal. Consider the case that $Y \sim N(\mu, \sigma^2)$, and $w(y) = (e^y + 1)/(e^y + 4)$ for $y \in \mathbb{R}$. Then the natural parameter is $\theta = (\theta_1, \theta_2)$ where $\theta_1 = \mu/\sigma^2$ and $\theta_2 = 1/\sigma^2$. The κ^o is 0.429 for $N(0,1)$ and 0.475 for $N(0,4)$. When N is unknown the maximum likelihood estimator $\hat{\theta}(w)$ of θ given w is the solution of equations:

$$\begin{aligned} \mu &= \bar{z}(t) - \sigma^2\Big[\frac{n}{m+n} - \frac{\alpha}{m+n}\frac{1}{\kappa(w,\theta)}\Big]\frac{\kappa'_\mu(w,\theta)}{\kappa(w,\theta)}, \\ \sigma^2 &= s^2(\mu,t)\Big\{1 + 2\sigma^2\Big[\frac{n}{m+n} - \frac{\alpha}{m+n}\frac{1}{\kappa(w,\theta)}\Big]\frac{\kappa'_{\sigma^2}(w,\theta)}{\kappa(w,\theta)}\Big\}^{-1}, \end{aligned}$$

where κ'_μ and κ'_{σ^2} are partial derivatives of κ with respect to μ and σ^2 and

$$\bar{z}(t) = \frac{\sum x_i + \sum y_j}{m+n}, \qquad s^2(\mu,t) = \frac{1}{m+n}\Big[\sum_{i=1}^{n}(x_i-\mu)^2 + \sum_{j=1}^{m}(y_j-\mu)^2\Big]. \quad (14)$$

These equations can be used to define an iterative algorithm for calculating $\hat{\theta}$, i.e. a little MM algorithm in the inner loop. When N is known, $\hat{\theta}$ is the solution of equations:

$$\begin{aligned} \mu &= \bar{z}(t) - \sigma^2\Big(\frac{N-n}{m+n}\Big)\Big[\frac{\kappa'_\mu(w,\theta)}{1-\kappa(w,\theta)}\Big], \\ \sigma^2 &= s^2(\mu,t)\Big/\Big\{1+2\sigma^2\Big(\frac{N-n}{m+n}\Big)\Big[\frac{\kappa'_{\sigma^2}(w,\theta)}{\kappa(w,\theta)}\Big]\Big\}, \end{aligned}$$

Table 4. The summary statistics under the Normal(0,4), $\kappa^o = 0.475$

	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
	Unknown N						Known N					
	$m = n/2$			$m = 0$			$m = n/2$			$m = 0$		
	$n\alpha_n = 0.25$			$\alpha_n = 0.037n^{-0.5}$			$n = 20$					
N	41.00	41.88	6.86	42.0	42.	6.890	41.00	41.88	6.859	42.00	42.	6.891
$\bar{t}$	0.541	0.545	0.38	0.83	0.82	0.440	0.541	0.545	0.380	0.827	0.82	0.441
$s(t)$	1.960	1.974	0.26	1.9	1.9	0.315	1.960	1.974	0.257	1.940	1.9	0.316
$\hat{\mu}$	-0.061	-0.060	0.42	-0.007	-0.007	0.488	-0.057	-0.053	0.417	-0.025	-0.028	0.491
$\hat{\sigma}$	1.900	1.907	0.26	1.8	1.8	0.311	1.930	1.929	0.255	1.850	1.9	0.317
d	0.311	0.323	0.11	0.32	0.33	0.101	0.290	0.321	0.083	0.290	0.32	0.084
$\hat{\kappa}$	0.446	0.475	0.12	0.38	0.39	0.054	0.456	0.461	0.075	0.453	0.46	0.077
	$n\alpha_n = 0.26$						$n = 50$					
N	105.0	105.5	11.05	0.01	110.	11.16	105.0	105.5	11.05	105.0	100.	11.16
$\bar{t}$	0.566	0.549	0.23	0.82	0.82	0.284	0.566	0.549	0.232	0.816	0.82	0.284
$s(t)$	2.010	2.009	0.17	2.0	2.0	0.198	2.010	2.009	0.166	1.970	2.0	0.198
$\hat{\mu}$	0.035	0.032	0.27	0.12	0.12	0.318	-0.043	-0.041	0.267	-0.003	0.002	0.319
$\hat{\sigma}$	1.940	1.948	0.17	1.9	1.9	0.205	1.950	1.947	0.170	1.870	1.9	0.199
d	0.304	0.320	0.11	0.31	0.32	0.103	0.255	0.288	0.056	0.250	0.28	0.053
$\hat{\kappa}$	0.484	0.502	0.11	0.46	0.46	0.062	0.461	0.464	0.049	0.461	0.46	0.050
	$n\alpha_n = 0.28$						$n = 100$					
N	210.0	210.2	15.27	210	210.	15.71	210.0	210.2	15.27	211.0	210.	15.71
$\bar{t}$	0.536	0.545	0.16	0.83	0.82	0.205	0.536	0.545	0.165	0.829	0.82	0.205
$s(t)$	2.030	2.027	0.12	2.0	2.0	0.142	2.030	2.027	0.117	1.980	2.0	0.142
$\hat{\mu}$	0.059	0.066	0.19	0.21	0.22	0.234	-0.042	-0.035	0.189	0.051	0.054	0.234
$\hat{\sigma}$	1.950	1.951	0.12	1.9	1.9	0.145	1.950	1.948	0.117	1.850	1.9	0.141
d	0.305	0.322	0.11	0.32	0.33	0.098	0.250	0.270	0.039	0.250	0.27	0.035
$\hat{\kappa}$	0.470	0.487	0.10	0.48	0.48	0.049	0.466	0.468	0.035	0.462	0.47	0.035

Notes. The case $m = 0$ is the worst scenario.

which can also be used to define an iterative procedure. Given θ, calculating $\hat{w}$ is again straightforward from (5) or (11). The equation (12) has a unique

solution $c \in (0,1)$ for reasonable α. However, there are many small values of c for which the two sides of (12) are fairly close for some α – false solutions. So, it is important that the precision value given to this equation is very small. To have an automatic algorithm to ensure that $\hat{\kappa}$ is the largest solution of the $c \in (0,1)$ in (12), we use the Golden section method (towards larger values) to find a solution of (12) and then 4 further random searches beyond the first solution. Hence, if the same random number generator is used in both generating the data and random searches, the data are slightly different in known N (no need for random searches) and unknown N cases, even if the same random seed is used. This presents no barrier in assessing performance of our estimates in the finite sample situation. The simulation results are presented in Table 4. With unknown N, the average iteration number used in the MM algorithm is $4.9, 7.5, 10.4$ respectively for $n = 20, 50, 100$ when $(\mu, \sigma) = (0,1)$, and $3.7, 4.1, 5.8$ when $(\mu, \sigma) = (0,2)$. Again, the iteration number is smaller when N is treated known. All the estimates get closer to the true values as n increases.

Remark. The distribution of $\hat{\kappa}$, based on $\alpha_n = 0.03n^{-1/2}$ for the case of unknown N and model Poisson(2), is skewed (for all n) and bimodal (for $n = 50, 100$), while those in Poisson(10), N(0,1) and $N(0,4)$ are quite symmetric and unimodal. This may be related to the fact that there are few *distinct* data points from Poisson(2) and hence κ is harder to estimate. So, in the case of Poisson(2), the median is a better measurement than the mean in terms of the goodness fit of $\hat{\kappa}$ to κ^o.

5. The MM Algorithm

The iterative algorithm in (8) (with different definitions for $\hat{\theta}$ and $\hat{w}$ in the two cases, N is known or unknown) is called the MM (maximization-maximization) algorithm, following the convention for the EM algorithm. The convergence of the iterates $(\hat{w}^k, \hat{\theta}^k), k \geq 1$, to the (penalized) maximum likelihood estimators may be deduced from the following simple results which have other applications. They are similar to Wu (1983) in some ways, but not in others – notably, the maximizing values may be boundary points.

Assumptions. Let $\mathcal{X}$ and $\mathcal{Y}$ denote metric spaces and let $\rho : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ be a function which is bounded below, not identically $+\infty$, and continuous (with respect to the topology of $\mathbb{R} \cup \{\infty\}$). Suppose that

$$\inf_{y'} \rho(x, y') < \infty \qquad \text{and} \qquad \inf_{x'} \rho(x', y) < \infty, \qquad \text{for all } x, y,$$

and that these infima are attained. Then the sets

$$\Psi(x) = \{y \in \mathcal{Y} : \ \rho(x, y) = \inf_{y'} \rho(x, y')\}$$

and

$$\Phi(y) = \{x \in \mathcal{X} : \ \rho(x,y) = \inf_{x'} \rho(x',y)\}$$

are non empty for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Of course, $\Psi(x)$ and $\Phi(y)$ are closed subsets of $\mathbb{R}$ for all x and y.

Lemma 3. *The Ψ and Φ enjoy the following continuity properties: suppose that $x_k \to x \in \mathcal{X}$ and $y_k \to y \in \mathcal{Y}$ as $k \to \infty$;*
(i) *if $y_k \in \Psi(x_k)$, for all $k \geq 1$, then $y \in \Psi(x)$;*
(ii) *if $x_k \in \Phi(y_k)$, for all $k \geq 1$, then $x \in \Phi(y)$.*

Proof. If $y_k \in \Psi(x_k)$, for all $k \geq 1$, then

$$\begin{aligned}\rho(x,y) &= \lim_{k\to\infty} \rho(x_k,y_k) = \lim_{k\to\infty} \inf_{y'} \rho(x_k,y') \\ &\leq \inf_{y'} \lim_{k\to\infty} \rho(x_k,y') = \inf_{y'} \rho(x,y'),\end{aligned}$$

so that $y \in \Psi(x)$. This proves (i). The proof of (ii) is similar.

The Algorithm. Given an initial point $x_0 \in \mathcal{X}$, let $y_0 \in \Psi(x_0)$,

$$x_k \in \Phi(y_{k-1}) \qquad \text{and} \qquad y_k \in \Psi(x_k), \qquad \text{for all } k \geq 1.$$

Then the following properties hold:

Property 1. *The sequence $\rho(x_k,y_k), k \geq 0$, is non-increasing, since*

$$\rho(x_k,y_k) \leq \rho(x_k,y_{k-1}) \leq \rho(x_{k-1},y_{k-1}), \qquad \textit{for all } k \geq 1.$$

So,

$$\rho_\# = \lim_{k\to\infty} \rho(x_k,y_k) = \lim_{k\to\infty} \rho(x_k,y_{k-1}) \quad \textit{exists.}$$

Property 2. *If $\{(x,y) : \rho(x,y) \leq c\}$ is compact for some $c > \inf_{x,y} \rho(x,y)$ and if $\rho(x_0,y_0) \leq c$, then the sequence (x_k,y_k), $k \geq 1$, is precompact (so that every subsequence contains a convergence subsequence).*

Property 3. *If $(x^\#,y^\#)$ is any limit point of the sequence (x_k,y_k), $k \geq 0$, then*

$$\rho(x^\#,y^\#) = \rho_\#, \quad y^\# \in \Psi(x^\#) \qquad \textit{and } x^\# \in \Phi(y^\#).$$

Proof. The first two assertions are clear from continuity of ρ and the continuity properties of Φ and Ψ in Lemma 3. For the third, let $\mathcal{K} \subseteq \{1,2,\ldots\}$ be a subsequence for which $(x_k,y_k) \to (x^\#,y^\#)$ as $k \to \infty$ through $\mathcal{K}$. Then

$$\begin{aligned}\rho(x^\#,y^\#) = \rho_\# &= \lim_{k\in\mathcal{K}} \rho(x_{k+1},y_k) = \lim_{k\in\mathcal{K}} \inf_{x'\in\mathcal{X}} \rho(x',y_k) \\ &\leq \inf_{x'\in\mathcal{X}} \lim_{k\in\mathcal{K}} \rho(x',y_k) = \inf_{x'} \rho(x',y^\#),\end{aligned}$$

so that $x^{\#} \in \Phi(y^{\#})$.

For the next three properties suppose $\mathcal{X}$ and $\mathcal{Y}$ are convex subsets of Euclidean spaces $\mathbb{R}^p$ and $\mathbb{R}^q$, and that ρ is differentiable on the set $Z = \{(x, y) : \rho(x, y) < \infty\}$.

Property 4. *Any limit point* $(x^{\#}, y^{\#})$ *of* $(x_k, y_k), k \geq 1$, *satisfies*

$$\frac{\partial \rho}{\partial x}(x^{\#}, y^{\#})'(x - x^{\#}) \geq 0, \quad \textit{for all } x \in \mathcal{X}, \tag{15}$$

and

$$\frac{\partial \rho}{\partial y}(x^{\#}, y^{\#})'(y - y^{\#}) \geq 0, \quad \text{for all } y \in \mathcal{Y}. \tag{16}$$

These are obvious necessary conditions for $y^{\#} \in \Psi(x^{\#})$ and $x^{\#} \in \Phi(y^{\#})$. For example, if $y \in \mathcal{Y}$ and $y^{\epsilon} = \epsilon y + (1 - \epsilon)y^{\#}$ for $0 < \epsilon < 1$, then $(x^{\#}, y^{\epsilon}) \in Z$ for sufficiently small ϵ. For such $\epsilon, \rho(x^{\#}, y^{\#}) \leq \rho(x^{\#}, y^{\epsilon})$ and, therefore,

$$0 \leq \lim_{\epsilon \searrow 0} \frac{\rho(x^{\#}, y^{\epsilon}) - \rho(x^{\#}, y^{\#})}{\epsilon} = \frac{\partial \rho}{\partial y}(x^{\#}, y^{\#})'(y - y^{\#}).$$

The term "convex function" below is used in the following extended sense: if $B \subseteq \mathcal{X} \times \mathcal{Y}$ is a (not necessarily convex) set, then a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ is said to be convex on B iff $f(\alpha w + \beta z) \leq \alpha f(w) + \beta f(z)$ whenever $w, z \in B$, $\alpha, \beta \geq 0$, and $\alpha + \beta = 1$.

Property 5. *If, in addition (to the conditions of Property* 4*), there is a closed set* $B \subseteq Z$ *for which* ρ *is convex on* B *and*

$$\{(x, y) : y \in \Psi(x)\} \subseteq B, \tag{17}$$

then ρ *is minimized at any limit point* $(x^{\#}, y^{\#})$ *of* $(x_k, y_k), k \geq 1$.

Proof. Clearly, $(x^{\#}, y^{\#}) \in B$ by Property 3 and (17); and $\inf_{x,y} \rho(x, y) = \inf_{(x,y) \in B} \rho(x, y)$ by (17). So, it suffices to show that $\rho(x^{\#}, y^{\#}) \leq \rho(x, y)$ for all $(x, y) \in B$. This follows from (15), (16) and the assumed convexity. For if $(x, y) \in B$, $x^{\epsilon} = \epsilon x + (1 - \epsilon)x^{\#}$, and $y^{\epsilon} = \epsilon y + (1 - \epsilon)y^{\#}$, $0 < \epsilon < 1$, then

$$\begin{aligned} \rho(x, y) - \rho(x^{\#}, y^{\#}) &\geq \frac{\rho(x^{\epsilon}, y^{\epsilon}) - \rho(x^{\#}, y^{\#})}{\epsilon}, \quad \text{for all } 0 < \epsilon < 1, \\ &\to \nabla \rho(x^{\#}, y^{\#})' \begin{pmatrix} x - x^{\#} \\ y - y^{\#} \end{pmatrix} \geq 0, \quad \text{as } \epsilon \searrow 0. \end{aligned}$$

Property 6. *If the conditions in Property* 2 *and* 5 *hold and if* ρ *attains its global minimum at a unique point* (x^0, y^0), *say, and* $\rho(x_0, y_0) \leq c$, *then* $(x_k, y_k) \to (x^0, y^0)$ *as* $k \to \infty$.

This is clear.

Application to the Maximum Likelihood Estimators. To apply this result to the log likelihood functions of Section 2, write $w_i = e^{\xi_i}, i = 1, \ldots, r$, and

$$\rho(\xi, \theta) = -l(w, \theta) \quad \text{and} \quad \rho^*(\xi, \theta) = -l^*_\alpha(w, \theta),$$

for $\xi \in \Xi = \{\xi \in \mathbb{R}^r : -\infty < \xi_1 \leq \cdots \leq \xi_r \leq 0\}$ and $\theta \in \Omega$. Then Ξ and Ω are convex, and ρ and ρ^* are bounded below and continuously differentiable. So, it remains to verify the compactness and convexity conditions in Properties 2 and 5.

Known N. For the case of known N, the compactness condition in Property 2 is clear, since $\sup_w l(w, \theta) \leq (m + n)[\theta' \bar{t} - \psi(\theta)]$, which approaches $-\infty$ as $||\theta|| \to \infty$ or θ approaches a boundary point of Ω, and $\sup_{\theta \in K} l(w, \theta) \to -\infty$ as $w_1 \searrow 0$ for any compact $K \subseteq \Omega$. It is shown below that the convexity condition in Property 5 is satisfied with

$$B = \Big\{(\xi, \theta) : \ \kappa(w, \theta) \leq \frac{n + m}{N + m}\Big\},$$

which is a closed but not necessarily convex subset of $\Xi \times \Omega$.

First, observe that $[\hat{w}(\theta), \theta] \in B$, for all $\theta \in \Omega$, by (6). Next, write $\phi(\xi, \theta) = \psi_w(\theta)$ and let S_i be the indicator of $[x_i, x_{i+1}), i = 1, \ldots, r$. Then ,

$$\phi(\xi, \theta) = \log\Big\{\int_{[x_1, \infty)} \exp(\theta' T + \xi' S) d\Lambda\Big\},$$

which is (jointly) convex in (ξ, θ). In fact, ϕ is strictly convex w.p.1. Now, simple algebra yields,

$$\rho(\xi, \theta) = -\sum_{i=1}^{r} n_i \xi_i - (m + n)\theta' \bar{t} + g[\phi(\xi, \theta), \psi(\theta)], \qquad \text{for all } \xi \in \Xi, \theta \in \Omega,$$

where

$$g(x, y) = -(N - n)\log[e^y - e^x] + (N + m)y = -(N - n)\log[1 - e^{x-y}] + (m + n)y$$

for all $-\infty < x < y < \infty$. It is easily seen that g is convex on $-\infty < x < y < \infty$ and that g is non-decreasing in each variable on the set

$$A = \{(x, y) \in \mathbb{R}^2 : x - y \leq \log(\frac{n + m}{N + m})\};$$

that is, if $(x_1, y_1), (x_2, y_2) \in A, x_1 \leq x_2, y_1 \leq y_2$, then $g(x_1, y_1) \leq g(x_2, y_2)$. Next observe that $[\phi(\xi, \theta), \psi(\theta)] \in A$ whenever $(\xi, \theta) \in B$. It follows easily that $g[\phi(\xi, \theta), \psi(\theta)]$ is convex in $(\xi, \theta) \in B$. To see this, let $(\xi^1, \theta^1), (\xi^2, \theta^2) \in$

$B, \alpha_1, \alpha_2 \geq 0, \alpha_1 + \alpha_2 = 1$, and $(\xi, \theta) = \alpha_1(\xi^1, \theta^1) + \alpha_2(\xi^2, \theta^2)$. Then, with obvious notational conventions,

$$\begin{aligned} x \equiv \phi(\xi, \theta) &\leq \alpha_1\phi(\xi^1, \theta^1) + \alpha_2\phi(\xi^2, \theta^2) = \alpha_1 x_1 + \alpha_2 x_2, \\ y \equiv \psi(\theta) &\leq \alpha_1\psi(\theta^1) + \alpha_2\psi(\theta^2) = \alpha_1 y_1 + \alpha_2 y_2 \end{aligned}$$

by convexity of ϕ and ψ, and

$$g(x, y) \leq g(\alpha_1 x_1 + \alpha_2 x_2, \alpha_1 y_1 + \alpha_2 y_2) \leq \alpha_1 g(x_1, y_1) + \alpha_2 g(x_2, y_2),$$

since g is non-decreasing on A. Therefore, the iterates $(\hat{w}^k, \hat{\theta}^k)$ defined by (5), (7) and (8) are precompact and any limit point is a maximum likelihood estimator of w, θ, using Properties 1-5.

Unknown N. Suppose that $\epsilon \geq \alpha > 0$. Then the compactness condition in Property 2 is clear for unknown N. For in this case Ξ is replaced by $\Xi = \{\xi \in \mathbb{R}^r : \log(\epsilon) \leq \xi_1 \leq \cdots \leq \xi_{r-1} \leq \xi_r = 0\}$, and $\psi(\theta) - \theta'\bar{t} \to \infty$ as $||\theta|| \to \infty$ or θ approaches a boundary point of Ω. For the convexity, note that

$$\rho^*(\xi, \theta) = -\sum_{i=1}^{r} n_i\xi_i + n[\phi(\xi, \theta) - \theta'\bar{t}] + m[\psi(\theta) - \theta'\bar{t}] + \frac{\alpha n}{\kappa},$$

where $\kappa = \exp[\phi(\xi, \theta) - \psi(\theta)]$. Taking its gradient with respect to (ξ, θ), we have

$$\nabla\rho^* = n\nabla\phi + m\nabla\psi - \alpha n\left[\frac{\nabla\phi - \nabla\psi}{\kappa}\right] + C,$$

where C is a constant with respect to ξ and θ, and

$$\begin{aligned} \nabla^2\rho^* &= n\nabla^2\phi + m\nabla^2\psi - \alpha n\left[\frac{\nabla^2\phi - \nabla^2\psi}{\kappa} - \frac{(\nabla\phi - \nabla\psi)(\nabla\phi - \nabla\psi)'}{\kappa}\right] \\ &= n\left(1 - \frac{\alpha}{\kappa}\right)\nabla^2\phi + \left(m + \frac{\alpha n}{\kappa}\right)\nabla^2\psi + \alpha n\left[\frac{(\nabla\phi - \nabla\psi)(\nabla\phi - \nabla\psi)'}{\kappa}\right] \quad (18) \end{aligned}$$

which is positive definite as $\kappa \geq \epsilon > \alpha$. In other words, ρ^* is strictly convex and has a unique minimum. Therefore, the iterates $(\hat{w}^k, \hat{\theta}^k)$ defined by (11), (13) and (8) approach the unique maximum likelihood estimator of (w, θ), as $k \to \infty$, using Properties 1-6.

6. Consistency

Identifiability. The family $\{f^*_{w,\theta} : w \in W, \theta \in \Omega\}$ is not identifiable, because $f^*_{w,\theta} = f^*_{cw,\theta}$ for any $w \in W, \theta \in \Omega$, and $c > 0$ for which $cw \leq 1$. A positive result is presented next.

Proposition 1. *Suppose that*

$$\limsup_{y\to\infty} |(\eta-\theta)'T(y)| = \infty, \qquad \text{for all } \eta \neq \theta. \tag{19}$$

If $u, w \in W$, $\eta, \theta \in \Omega$, *and* $f^*_{u,\eta} = f^*_{w,\theta}$ *a.e.* (Λ), *then* $\eta = \theta$ *and* $u = cw$ *a.e.* (Λ) *for some* $c > 0$.

Proof. If $u, w \in W$, $\eta, \theta \in \Omega$ and $f^*_{u,\eta} = f^*_{w,\theta}$ a.e. (Λ), then

$$u(y)\exp[\eta'T(y)] = cw(y)\exp[\theta'T(y)] \qquad \text{a.e. } y\ (\Lambda), \tag{20}$$

where $c = \kappa(u,\eta)\exp[\psi(\eta)-\psi(\theta)]/\kappa(w,\theta)$, a positive constant. If y is sufficiently large, then both sides of (20) are positive, and

$$\left|\log\left[\frac{u(y)}{w(y)}\right]\right| = |(\theta-\eta)'T(y) + \log(c)|. \tag{21}$$

The left side of (21) has a finite limit as $y \to \infty$, and the right side has an infinite limit superior if $\eta \neq \theta$ by (19). So, $\eta = \theta$. That $u = cw$ a.e. (Λ) then follows directly from (20).

Corollary 1. *If* $\kappa(u,\eta)=\kappa(w,\theta)$ (*in addition to the conditions of the proposition*), *then* $\eta = \theta$ *and* $u = w$.

For a one parameter exponential family ($p = 1$), condition (19) is satisfied, if $\limsup_{y\to\infty} |T(y)| = \infty$. In particular, (19) is satisfied if $T(y) = y$. For $p \geq 2$, sufficient conditions for (19) are that

$$\lim_{y\to\infty} |T_j(y)| = \infty \quad \text{and} \quad \lim_{y\to\infty} \frac{|T_j(y)|}{|T_{j+1}(y)|} = 0, \qquad \text{for all } j = 1, \ldots, p-1. \tag{22}$$

For example, suppose that (22) holds with $p = 2$. Then $|(\eta-\theta)'T(y)| = |(\eta_1 - \theta_1)T_1(y) + (\eta_2-\theta_2)T_2(y)| \to \infty$ if either $\eta_2 \neq \theta_2$ or $\eta_2 = \theta_2$ and $\eta_1 \neq \theta_1$; that is, if $\eta \neq \theta$. It is easily verified that (22) is satisfied by the two parameter normal family, with $T(y) = (y, -y^2/2)', -\infty < y < \infty$, and the two parameter gamma family, with $T(y) = [\log(y), -y]', 0 < y < \infty$.

Distance in W. To discuss consistency, it is necessary to define distance in W. Two functions $v, w \in W$ are regarded as equivalent if $v = w$ a.e. (Λ). The same symbol w is used to denote a function and the equivalence class containing it. If G is any probability distribution which has the same null sets as Λ, then

$$d(v,w) = d_G(v,w) = \int_{\mathcal{Y}} |v-w| dG, \qquad \text{for all } v, w \in W,$$

defines a metric for W. It is easily seen that all such metrics generate the same topology and that κ is a continuous function on the product space $W \times \Omega$.

The same symbol H is used for both a probability distribution on $\mathcal{Y}$ and its distribution function in the next lemma.

Lemma 4. *Let H be a probability distribution for which $H << \Lambda$; and let $H_k, k \geq 1$, be probability distributions for which $\sup_{y\in\mathcal{Y}} |H_k(y) - H(y)| \to 0$, as $k \to \infty$. Further, let $\delta > 0$, and $w_1, w_2, \ldots \in W$ be functions for which $\int_{\mathcal{Y}} w_k dH_k \geq \delta$ a.e. for k (all but a finite number). If $\mathcal{K} \subseteq \{1, 2, \ldots\}$, then there is a subsequence $\mathcal{K}_0 \subseteq \mathcal{K}$ and a $w \in W$ for which $\int_{\mathcal{Y}} w dH \geq \delta$ and $w_k \to w$, as $k \to \infty$ through $\mathcal{K}_0$. In particular, $\{w \in W : \int_{\mathcal{Y}} w dH \geq \delta\}$ is compact for any $\delta > 0$ and any $H << \Lambda$.*

Proof. By a simple diagonalization argument, there is a subsequence $\mathcal{K}_0 \subseteq \mathcal{K}$ and a $w \in W \cup \{0\}$ for which $w_k(x) \to w(x)$ as $k \to \infty$ through $\mathcal{K}_0$, at all continuity points x of w and at all atoms x of Λ. Then

$$\int_{\mathcal{Y}} w_k dH_k - \int_{\mathcal{Y}} w dH = \int_{\mathcal{Y}} (H - H_k) dw_k + \int_{\mathcal{Y}} (w_k - w) dH,$$

which approaches zero as $k \to \infty$ through $\mathcal{K}_0$, by the assumed uniform convergence of H_k to H and the Bounded Convergence Theorem. It follows that $\int_{\mathcal{Y}} w dH \geq \delta$ and, therefore, that $w \in W$. Another application of the Bounded Convergence Theorem shows that $d_G(w, w_k) \to 0$, as $k \to \infty$ through $\mathcal{K}_0$, for any probability distribution $G << \Lambda$; that is, $w_k \to w$, as $k \to \infty$ through $\mathcal{K}_0$. This establishes the first assertion of the lemma, and the second follows by specializing the first to $H_k = H$, for all $k \geq 1$.

Consistency. To simplify the notation, denote the true values of w and θ by w^o and θ^o, let $\kappa^o = \kappa(w^o, \theta^o)$, and write f and f^* for f_{θ^o} and $f^*_{w^o,\theta^o}$ respectively. It is convenient (and efficient) to suppose that there are three independent sequences of random variables $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, and $n = n_N, N \geq 1$, defined on an appropriate probability space, for which $X_1, X_2, \ldots \sim f^*$ are i.i.d., $Y_1, Y_2, \ldots \sim f$ are i.i.d., and $n \sim \text{Binomial}(N, \kappa^o)$ for all $N \geq 1$. (The dependence of n on N is suppressed in the notation.) For the limiting operations $N \to \infty$ and $m = m_N$ depends on N in such a manner that $m/N \to \gamma\kappa^o$, where $\gamma \geq 0$. Of course, $n/N \to \kappa^o$ w.p.1.

If $n, X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are observed, then the log likelihood function and conditional (penalized) log likelihood function given n may be written as

$$\begin{aligned} l_{m,n}(w, \theta) &= R_n(w) + S_{m,n}(\theta) + (N - n) \log[1 - \kappa(w, \theta)], \\ l^*_{\alpha,m,n}(w, \theta) &= R_n(w) + S_{m,n}(\theta) - n \log[\kappa(w, \theta)] - \frac{\alpha n}{\kappa(w, \theta)}, \end{aligned}$$

where

$$R_n(w) = \sum_{i=1}^{n} \log[w(X_i)]$$

and

$$S_{m,n}(\theta) = \sum_{i=1}^{n} \theta' T(X_i) + \sum_{j=1}^{m} \theta' T(Y_j) - (m+n)\psi(\theta),$$

for $w \in W$ and $\theta \in \Omega$. For $\epsilon \geq 0$, $w \in W, \theta \in \Omega$, and $n \geq 1$, let

$$R_{\epsilon,k}(w) = \sum_{i=1}^{k} \log[\epsilon \vee w(X_i)], \quad k \geq 1,$$
$$r_\epsilon(w) = \int_{\mathcal{Y}} \log[\epsilon \vee w(y)] f^*(y) \Lambda(dy),$$

and

$$s(\theta) = \theta' \nabla \psi_{w^o}(\theta^o) + \gamma \theta' \nabla \psi(\theta^o) - (1+\gamma)\psi(\theta).$$

Lemma 5. *For any compact $K \subseteq \Omega$,*

$$\sup_{\theta \in K} \left| \frac{S_{m,n}(\theta)}{N} - \kappa^o s(\theta) \right| \to 0, \quad w.p.1. \qquad as\ N \to \infty. \tag{23}$$

Further, if $m/N - \gamma\kappa_0 = O(N^{-1/2})$, as $N \to \infty$, and

$$K_n = \{\theta \in \Omega : \psi(\theta) \leq c \log n\},$$

where $0 < c < \infty$, then

$$\sup_{\theta \in K_n} \left| \frac{S_{m,n}(\theta)}{N} - \kappa^o s(\theta) \right| = O\left(\frac{\log^2 n}{\sqrt{n}} \right), \quad w.p.1. \qquad as\ N \to \infty. \tag{24}$$

Proof. The first statement in (23) follows easily from the law of large numbers, applied to $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ The second in (24) is nearly as transparent from the law of the iterated logarithm and the relation $||\theta|| \leq O[\psi(\theta)]$ as $||\theta|| \to \infty$, using Lemma 2.

Lemma 6. *For all $0 < \epsilon < 1$,*

$$\sup_{w \in W} \left| \frac{1}{k} R_{\epsilon,k}(w) - r_\epsilon(w) \right| \to 0, \quad w.p.1. \qquad as\ k \to \infty.$$

Furthermore, if $\epsilon = \epsilon_k \to 0$ as $k \to \infty$ and $\log \epsilon_k^{-1} = O(\log k)$, then

$$\sup_{w \in W} \left| \frac{1}{k} R_{\epsilon,k}(w) - r_\epsilon(w) \right| = O(\frac{\log^2 k}{\sqrt{k}}), \quad w.p.1. \qquad as\ k \to \infty.$$

Proof. Let F^* denote the distribution function of X_1 and let F_k be the empirical distribution function of $X_1, \ldots, X_k$. Then

$$\left|\frac{1}{k}R_{\epsilon,k}(w) - r_\epsilon(w)\right| = |\int_{-\infty}^{\infty} \log(\epsilon \vee w) d(F_k - F^*)|$$
$$= |\int_{-\infty}^{\infty} (F_k - F^*) d\log(\epsilon \vee w)| \le \sup_x |F_k(x) - F^*(x)| \cdot \log(\frac{1}{\epsilon})$$

which goes to zero w.p.1. for fixed $\epsilon > 0$ as $k \to \infty$, using the Glivenko-Cantelli theorem. The second assertion follows from the law of the iterated logarithm for $\sup_x |F_k(x) - F^*(x)|$ if F^* is continuous. (See, for example, Csörgő and Revesz (1981), pp 157.) If F^* is discrete and the probability space is sufficiently rich, then $\sup_x |F_k(x) - F^*(x)| \le \sup_u |G_k(u) - G(u)|$ for a uniform distribution G and some uniform empirical distributions G_k, and we can apply the iterated logarithm to $\sup_u |G_k(u) - G(u)|$.

Known N. For the first theorem, let $\hat{w}_N$ and $\hat{\theta}_N$ denote approximate maximum likelihood estimators, so that

$$l_{m,n}(\hat{w}_N, \hat{\theta}_N) \ge \sup_{w \in W, \theta \in \Omega} l_{m,n}(w, \theta) - \delta_N \quad \text{w.p.1.}, \tag{25}$$

where $\delta_1, \delta_2, \ldots$ is a sequence of real numbers for which $\delta = o(N)$.

Theorem 1. *If the conditions* (19) *and* (25) *are satisfied, then*

$$\lim_{N \to \infty} [\hat{w}_N, \hat{\theta}_N] = (w^o, \theta^o) \quad w.p.1.$$

Proof. It is first shown that there are compact $U \subseteq W$ and $K \subseteq \Omega$ for which

$$(\hat{w}_N, \hat{\theta}_N) \in U \times K \quad \text{and} \quad l_{m,n}(\hat{w}_N, \hat{\theta}_N) \ge \sup_{w \in U, \theta \in K} l_{m,n}(w, \theta) - \delta_N \quad \text{w.p.1.} \tag{26}$$

a.e. for N (all but a finite number). To see this first observe that $l_{m,n}(\hat{w}_N, \hat{\theta}_N)/N \ge l_{m,n}(w^o, \theta^o)/N + o(1) \to \kappa^o r_0(w^o) + \kappa^o s(\theta^o) + (1 - \kappa^o)\log(1 - \kappa^o) > -\infty$ w.p.1. as $N \to \infty$. So, there is a $1 \le C < \infty$ for which

$$l_{m,n}(\hat{w}_N, \hat{\theta}_N) \ge -CN \tag{27}$$

a.e. for N. Next, by Lemma 1, there is a compact $J \subseteq \nabla\psi(\Omega)$ for which $\bar{t}_N = [T(X_1) + \cdots + T(Y_m)]/(m + n) \in J$ a.e. for N w.p.1. By Lemma 2, there is a compact $K \subseteq \Omega$ for which $S_{m,n}(\theta)/N \le -2C$ for all $\theta \notin K$ a.e. for N. Finally,

$$l_{m,n}(w, \theta) = R_n(w) + S_{m,n}(\theta) + (N - n)\log[1 - \kappa(w, \theta)]$$
$$\le R_n(w) + S_{m,n}(\theta) \le S_{m,n}(\theta) \tag{28}$$

for all $w \in W$ and $\theta \in \Omega$. Since the right side of (28) is less than the right side of (27) for $\theta \notin K$ a.e. for N w.p.1., it follows that $\hat{\theta}_N \in K$ a.e. for N, w.p.1. Next, letting $B = \kappa^o \sup_{\theta \in K} |s(\theta)| + 1$, it follows that $-CN \leq l_{m,n}(\hat{w}_N, \hat{\theta}_N) \leq R_n(\hat{w}_N) + BN$ a.e. for N. So, by Lemma 6 and Jensen's Inequality,

$$-(B + C) \leq \frac{n}{N} \frac{1}{n} R_n(\hat{w}_N) \leq \frac{\kappa^o}{2} \log(\int_{\mathcal{Y}} \hat{w}_N dF_n)$$

a.e. for N, w.p.1., where F_n denotes the empirical distribution function. That is, the sequences $\hat{w}_N, N \geq 1$, and F_n satisfy the conditions of Lemma 4 with $\delta = \exp[-2(B + C)/\kappa_0]$, w.p.1. It follows that $\hat{w}_N$ is in the compact set $U = \{w \in W : \kappa(w, \theta^o) \geq \delta/2\}$ a.e. for N. Relation (26) follows.

By (26), $(\hat{w}_N, \hat{\theta}_N)$ is relatively compact w.p.1. It is also clear that for all $\epsilon > 0$ and N a.e.

$$\begin{aligned} \frac{1}{N} l_{m,n}(\hat{w}_N, \hat{\theta}_N) &\leq \sup_{w \in U, \theta \in K} \frac{1}{N} \{R_{\epsilon,n}(w) + S_{m,n}(\theta) + (N - n) \log[1 - \kappa(w, \theta) + \epsilon], \\ &\to \sup_{w \in U, \theta \in K} \{\kappa^o r_\epsilon(w) + \kappa^o s(\theta) + (1 - \kappa^o)[1 - \kappa(w, \theta) + \epsilon]\} \end{aligned}$$

w.p.1. as $N \to \infty$; and the right side approaches $\sup_{w \in U, \theta \in K} \{\kappa^o r_0(w) + \kappa^o s(\theta) + (1 - \kappa^o)[1 - \kappa(w, \theta)]\}$ as $\epsilon \searrow 0$, by Dini's Theorem. If $(\hat{w}, \hat{\theta})$ denotes any limit point (possibly random) of the sequence $(\hat{w}_N, \hat{\theta}_N), N \geq 1$, and $\hat{\kappa} = \kappa(\hat{w}, \hat{\theta})$, then by the dominated convergence theorem,

$$\kappa^o \{r_0(\hat{w}) - r_0(w^o)\} + (1 - \kappa^o) \log[\frac{1 - \hat{\kappa}}{1 - \kappa^o}] + \kappa^o [s(\hat{\theta}) - s(\theta^o)] \geq 0.$$

Letting $\hat{f}^* = f^*_{\hat{w}, \hat{\theta}}, \hat{f} = f_{\hat{w}, \hat{\theta}}$ and $\hat{\kappa} = \kappa(\hat{w}, \hat{\theta})$, this inequality may be rewritten

$$\kappa^o \int_{\mathcal{Y}} \log(\frac{\hat{f}^*}{f^*}) f^* d\Lambda + \gamma \kappa^o \int_{\mathcal{Y}} \log(\frac{\hat{f}}{f}) f d\Lambda + \kappa^o \log(\frac{\hat{\kappa}}{\kappa^o}) + (1 - \kappa^o) \log(\frac{1 - \hat{\kappa}}{1 - \kappa^o}) \geq 0.$$

Finally, the latter inequality requires $\hat{\kappa} = \kappa^o$ and $\hat{f}^* = f^*$ by Jensen's Inequality, and, therefore, $(\hat{w}, \hat{\theta}) = (w^o, \theta^o)$ by Proposition 1. Thus, (w^o, κ^o) is the unique limit point of $(\hat{w}_N, \hat{\theta}_N), N \geq 1$, w.p.1.

Unknown N. In the second theorem, let $\alpha = \alpha_n$, where $0 < \alpha_1, \alpha_2, \ldots, \epsilon_1, \epsilon_2, \ldots$ are sequences for which $\epsilon_k > \alpha_k$ for all $k \geq 1, \epsilon_k \to 0$, and $\alpha_k^{-1} = o(k^{1/2}/\log^2 k)$ as $k \to \infty$. Suppose also that $m/N = \gamma \kappa^o + O(N^{-1/2})$ as $N \to \infty$. As in the case of known N, let $\hat{w}_N$ and $\hat{\theta}_N$ denote approximate conditional penalized maximum likelihood estimators so that

$$l^*_{\alpha,m,n}(\hat{w}_N, \hat{\theta}_N) \geq \sup_{\epsilon_n \leq w \in W, \theta \in \Omega} l^*_{\alpha,m,n}(w, \theta) - \delta_N \quad \text{w.p.1., for all } N \geq 1, \qquad (29)$$

where $\delta_1, \delta_2, \ldots$ is a sequence of real numbers for which $\delta_N = O(N^{1/2} \log N)$.

Theorem 2. *If $w^o(\infty) = 1$, condition (19) and those in the previous paragraph are satisfied, then*

$$\lim_{N\to\infty} [\hat{w}_N, \hat{\theta}_N] = (w^o, \theta^o) \qquad w.p.1.$$

Proof. The proof is similar to that of Theorem 1, but the proof of the relative compactness is more complicated. In this proof, the following inequalities are needed. If J is any compact subset of $\nabla\psi(\Omega)$, then there are a $B = B_J > 0$ and a $\eta = \eta_J, 0 < \eta < 1$ for which

$$\theta' t \le B + (1-\eta)\, \psi(\theta), \qquad \text{for all } \theta \in \Omega, \quad t \in J.$$

This follows easily from Lemma 3.5 and Lemma 5.3 of Brown (1986), pp. 73-74 and pp. 146 as our Lemma 2 does. So, for $\epsilon > 0$ simple algebra yields

$$\begin{aligned} &\left\{(w,\theta) \in W \times \Omega : \ \kappa(w,\theta) \ge \epsilon,\ l^*_{\alpha,m,n}(w,\theta) \ge -CN\right\} \\ &\subseteq \left\{(w,\theta) \in W \times \Omega : \ \psi(\theta) \le \eta^{-1}\left[B + \frac{n}{m+n}\log\kappa^{-1} + \frac{CN}{m+n}\right]\right\}. \quad (30) \end{aligned}$$

The above inclusion also holds with $\log\kappa^{-1}$ replaced by $\log\epsilon^{-1}$, obviously.

Next, it is shown that there are compact $U_1, U_2, \ldots \subseteq W$ and $K_1, K_2, \ldots \subseteq \Omega$ for which $(\hat{w}_N, \hat{\theta}_N) \in U_n \times K_n$ for a.e. N, w.p.1. Let $U_k = \{w \in W : \ w \ge \epsilon_k\}$, $k = 1, 2, \ldots$ Then clearly $\hat{w}_N \in U_n$ for all N. As in the proof of Theorem 1, there are a compact $J \subseteq \Omega$ and a constant $0 < C < \infty$ for which $\bar{t}_N \in J$ and $l^*_{\alpha,m,n}(\hat{w}_N, \hat{\theta}_N) \ge -CN$ for a.e. N, w.p.1. Moreover, $N/n \le 2/\kappa^o$ for a.e. N, w.p.1. It follows from (30) that $(\hat{w}_N, \hat{\theta}_N) \in U_n \times K_n$ with

$$\begin{aligned} K_j &= \left\{\theta \in \Omega : \psi(\theta) \le \eta^{-1}\left[B + \log\kappa_j^{-1} + \frac{2C}{\kappa^o}\right]\right\} \\ &\subseteq \left\{\theta \in \Omega : \psi(\theta) \le \eta^{-1}\left[B + \log\epsilon_j^{-1} + \frac{2C}{\kappa^o}\right]\right\}. \end{aligned} \quad (31)$$

Observe that $r_\epsilon(w) = r_0(w)$ if $w \ge \epsilon$. So, by the second part of Lemma 5, Lemma 6, and the law of the iterated logarithm,

$$\begin{aligned} \frac{1}{N} l^*_{\alpha,m,n}(\hat{w}_N, \hat{\theta}_N) &= \kappa^o\{r_0(\hat{w}_N) + s(\hat{\theta}_N) - \log\hat{\kappa}_N - \frac{\alpha}{\hat{\kappa}_N}\} + O(\frac{\log^2 n}{\sqrt{n}}), \\ \frac{1}{N} l^*_{\alpha,m,n}(w^o \vee \epsilon_n, \theta^o) &\ge \kappa^o\{r_0(w^o) + s(\theta^o) - \log\kappa^o - \frac{\alpha}{\kappa^o}\} + O(\frac{\log^2 n}{\sqrt{n}}), \end{aligned}$$

w.p.1. as $N \to \infty$. Since $l^*_{\alpha,m,n}(\hat{w}_N, \hat{\theta}_N) \geq l^*_{\alpha,m,n}(w^o \vee \epsilon_n, \theta^o) - \delta_N$ and

$$r_0(\hat{w}_N) + s(\hat{\theta}_N) - \log \hat{\kappa}_N \leq r_0(w^o) + s(\theta^o) - \log \kappa^o \ ,$$

by the information inequality, as before, it follows from (29) that

$$\frac{\alpha}{\kappa^o} - \frac{\alpha}{\hat{\kappa}_N} \geq -\frac{\delta_N}{N} + O(\frac{\log^2 n}{\sqrt{n}}) = O(\frac{\log^2 n}{\sqrt{n}}), \qquad \text{w.p.1.}$$

and, therefore, that

$$\liminf_{N\to\infty} \hat{\kappa}_N \geq \kappa^o, \qquad \text{w.p.1.} \tag{32}$$

in view of the assumptions on α. It then follows easily from (31) and Lemma 4 that $(\hat{w}_N, \hat{\theta}_N), N \geq 1$, is relatively compact.

Letting $(\hat{w}, \hat{\theta})$ denote any limit point of $(\hat{w}_N, \hat{\theta}_N), N \geq 1$, and arguing as in the proof of Theorem 1, we find that

$$\kappa^o \int_{\mathcal{Y}} \log(\frac{\hat{f}^*}{f^*}) f^* d\Lambda + \gamma\kappa^o \int_{\mathcal{Y}} \log(\frac{\hat{f}}{f}) f d\Lambda \geq 0$$

and hence by Jensen's inequality,

$$\hat{\theta} = \theta^o, \qquad \text{and} \qquad \frac{\hat{w}}{\hat{\kappa}} = \frac{w^o}{\kappa^o}, \quad \text{a.e. } (\Lambda).$$

Clearly, $\hat{\kappa} \geq \kappa^o$ by (32); and $\hat{\kappa} \leq \kappa^o$ by the above equation, since $\hat{w}(\infty) \leq 1$ and $w^o(\infty) = 1$. That is, $\hat{\kappa} = \kappa^o, \hat{w} = w^o$, and $\hat{\theta} = \theta^o$. The theorem follows.

7. Concluding Remarks

Two important questions were left unanswered, asymptotic distributions of the estimators and the choice of the smoothing parameter α_n. These are related. Some possible approaches are described below.

Asymptotic Distributions. If θ were known, then the asymptotic distribution of $\hat{w}_n$ could be determined. For example, if θ is known, the distribution F_θ of f_θ is continuous, and $X \sim F^*_\theta$, then $Z = 1 - F_\theta(X)$ has density

$$g(z) = \frac{w[F^{-1}(1-z)]}{\kappa(w, \theta)},$$

and the problem becomes one of estimating a non-increasing density. This problem may be solved by using strong approximation, as in Groeneboom (1985), and $n^{1/3}[\hat{g}_n(z) - g(z)]$ has a (fairly complicated) limiting distribution, if g is sufficiently smooth near z. (See Woodroofe and Sun (1993) for details.) It seems reasonable to conjecture that similar results may be obtained for the unknown θ case

provided that there is enough good data; and examination of the proofs suggests that "enough" means that $n/m^{3/2} \to 0$. Return to the case of known θ and continuous F_θ; and let $l_n(w, \theta)$ denote the log-likelihood function of $Z_1, \ldots, Z_n \sim g$. It may be shown that

$$\sup_{w \in \mathcal{W}} [l_n(w, \theta) - l_n(1, \theta)] = O_p(\log n), \tag{33}$$

even without smoothing, by using results of Groeneboom and Pyke (1983) and equations like

$$\int_0^\infty \log \tilde{f}_n dF_n = \int_0^\infty \log \tilde{f}_n d\tilde{F}_n$$

of Woodroofe and Sun (1993), p510. Here $\tilde{f}_n, \tilde{F}_n$ are the nonparametric maximum likelihood estimates of the density and distribution under the monotone constraint and F_n is the empirical distribution function. Moreover, if an appropriate smoothing parameter is included in the log-likelihood, then an asymptotic distribution may be obtained for the left side of (33). Results like this could be used to construct likelihood ratio tests for the presence of a bias.

The Choice of α_n. The choice of α_n presents a thorny question when N is unknown. If α_n is too small, then $\hat{\kappa}$ has a negative bias and is highly variable, due largely to some very small values; and too large values of α_n lead to a substantial positive bias in $\hat{\kappa}$ and a flat $\hat{w}$. For estimating a non-increasing density, Sun and Woodroofe (1996) used the asymptotic distributions to show that the asymptotically optimal choice of α_n is of the form $cn^{-2/3}$, and they proposed some adaptive estimators of c. It is not clear that their conditions are satisfied in the present context, however, even if θ is known or m is large (so that θ may be estimated accurately). Some rough knowledge of κ may be necessary to choose c intelligently.

Knowledge of κ was used in the simulation studies reported in Section 4 in that α_n was chosen to make $\hat{\kappa}$ have a small bias.

An approach which does not require exact knowledge of κ is to use a jackknife (on κ) and/or least squares cross validation (on f^* and g). For example, if $\alpha_n = cn^{-0.5}$, then c might be chosen to minimize an expression like

$$\frac{1}{\hat{\kappa}_c(1-\hat{\kappa}_c)} \left\{ \frac{n-1}{n} \sum (\hat{\kappa}_{-i,c} - \hat{\kappa}_{\cdot,c})^2 + (n-1)^2 (\hat{\kappa}_c - \hat{\kappa}_{\cdot,c})^2 \right\},$$

where the subscript "$-i$" means that the ith data point has been omitted and the subscript "$\cdot$" denotes an average of $*_{-i}$'s, and rough knowledge of κ might be used to restrict the range of the search for c. The authors have conducted simulation studies of this alternative (not reported in detail here) for the Poisson and normal examples. It works in about 90% of the cases, and the results from

the cross validation usually have smaller variance than those from the jackknife. A general warning for cross validation/jackknife is that specifying the range of the search is important. (See, for example, Silverman (1986), p48ff.) Here, we searched c over $(0.01, 0.05)$ for $\kappa \leq 0.5$ and over $(0.06, 0.3)$ for $\kappa > 0.5$. The estimated c depends on the data (no longer a fixed value as in Tables 3 and 4). The distribution of the corresponding $\hat{\kappa}$ became unimodal for all cases we considered (cf. the remark at the end of Section 4).

Rough knowledge about the shape of w may also be useful for the choice of α_n. To see how, observe the following two artifacts of the estimators: it is always the case that $\hat{w}_n(x_n) = 1$ even if $\hat{w}_n(x_{n-1})$ is small; and it is often the case that $\hat{w}$ is flat if α_n is too large. If either of these features seems too pronounced, then the choice of α_n may be too large or too small. For example, for the moose data in Table 1, the choice $n\alpha_n = .4$ leads to an estimator for which $\hat{w}_n(6) = 1$, but $\hat{w}_n(5) = .5111$; and the choice $n\alpha_n = .8$ leads to $\hat{w}_n(2) \approx \hat{w}_n(6)$. Both of these conclusions seem questionable. Thus, it may be useful to examine the estimators for several values of $n\alpha_n$. This approach has been advocated by Silverman (1986) in a related context.

Acknowledgement

Research is supported in part by the National Science Foundation under grants DMS-92-03357 and DMS-95-04515. Thanks are due to the associate editor and referees for suggesting a more general model described at the end of Section 2 and helpful comments on an earlier draft of this paper.

A. Appendix: An Optimization Theorem

Let $r > r_0$ be two non-negative integers; and let $\Xi = \{\xi \in \mathbb{R}^r : -\infty < \xi_1 \leq \cdots \leq \xi_r < \infty\}$. Next let g be a concave function on Ξ with partial derivatives:

$$\frac{\partial g}{\partial \xi_k} = z_k - cp_k w_k \quad \text{for all } k = 1, \ldots, r, \tag{A.1}$$

where $c > 0, p_1, \ldots, p_r \geq 0$ and $z_1, \ldots, z_r$ are known constants, $w_k = h(\xi_k)$ and h is a known strictly increasing function. Finally let h^{-1} denote the inverse function of h and let

$$\tilde{w}_k = \max_{i \leq k} \min_{k \leq j \leq r - r_0} \frac{z_i + \cdots + z_j}{p_i + \ldots + p_j}, \quad \text{for all } k = 1, \ldots, r - r_0. \tag{A.2}$$

Then g is maximized on Ξ by $\tilde{\xi}$ where $\tilde{\xi}_k = h^{-1}(\tilde{w}_k)$ for $k = 1, \ldots, r$, where $r_0 = 0$. (See, for example, Theorem 1.4.4 of Robertson, Wright, and Dykstra (1988) (RWD thereafter).) We have the following extension theorem.

Theorem 3. *With the notations above, g is maximized in*

$$\Xi_{a,b} = \{\xi \in \Xi : a \le \xi_1 \le \cdots \le \xi_{r-r_0+1} = \cdots = \xi_r = b\},$$

by $\hat{\xi}$, where $\hat{\xi}_{r-r_0+1} = \hat{\xi}_r = b$ and $\hat{\xi}_k = h^{-1}(\hat{w}_k(c))$ for $k = 1, \ldots, r - r_0$, with

$$\hat{w}_k(c) = \max\{h(a), \min[\frac{\tilde{w}_k}{c},\ h(b)]\} \tag{A.3}$$

and $\tilde{w}_k$ in (A.2).

Moreover, if $c = H(w)$ in (A.1) *and the equation $c = H(\hat{w}(c))$ has a positive solution, say $c = \hat{c}$, then g is maximized by $\hat{\xi}$ where*

$$\hat{\xi}_k = h^{-1}(\hat{w}_k(\hat{c})) \quad \text{for } k = 1, \ldots, r - r_0, \quad \text{and} \quad \hat{\xi}_{r-r_0+1} = \hat{\xi}_r = b \tag{A.4}$$

with $\hat{w}_k(c)$ defined in (A.3).

Proof. We prove the more general result (A.4) directly and shall write $\hat{w}$ for $\hat{w}(\hat{c})$. As g is concave in ξ on $\Xi_{a,b}$, a necessary and sufficient condition for $\hat{\xi} \in \Xi_{a,b}$ to maximize g is that

$$\sum_{k=1}^{r-r_0} [z_k - H(\hat{w})p_k\hat{w}_k]\,(\xi_k - \hat{\xi}_k) \le 0, \qquad \text{for all } \xi \in \Xi_{a,b}. \tag{A.5}$$

By Theorems 1.3.2 and 1.3.6 of RWD, for $\tilde{w}$ defined in (A.2) and any real valued function Ψ

$$\sum_{k=1}^{r-r_0} [z_k - p_k\tilde{w}_k]\,(\xi_k - \Psi(\tilde{w}_k)) \le 0, \qquad \text{for all } -\infty < \xi_1 \le \cdots \le \xi_{r-r_0} < \infty. \tag{A.6}$$

Besides, if J is any set of the form $J = \{k : \tilde{w}_k = u\}$ for some $u \in \mathbb{R}$, then

$$\sum_{k\in J} p_k\tilde{w}_k = \sum_{k\in J} z_k. \tag{A.7}$$

(See, for example, Theorem 1.3.5 of RWD). Of course, (A.7) then holds for any J of the form $J = \{k : d_1 \le \tilde{w}_k \le d_2\}$ for $d_1, d_2 \in \mathbb{R}$.

It is shown next that the necessary and sufficient condition (A.5) is satisfied with the choice of $\hat{w}_k, k = 1, \ldots, r$, given in (A.4). Fix a $\xi \in \Xi_{a,b}$ and $c = \hat{c} = H(\hat{w})$ throughout the verification. Three cases are considered below.

Case (i). If $h(a) \le \tilde{w}_1/c \le \tilde{w}_{r-r_0}/c \le h(b)$, then $\hat{w}_k = \tilde{w}_k/c$ or $\hat{\xi}_k = h^{-1}(\tilde{w}_k/c)$ for $k = 1, \ldots, r - r_0$. So, the necessary and sufficient condition (A.5) follows easily from (A.6) with $\Psi(x) = h^{-1}(x/c)$.

Case (ii). Suppose that there are two integers $1 < t < s \le r - r_0$ for which

$$\tilde{w}_{t-1} < h(a)c, \quad \tilde{w}_t \ge h(a)c \qquad \text{and} \qquad \tilde{w}_{s-1} \le h(b)c, \quad \tilde{w}_s > h(b)c.$$

In this case there are three groups of $\hat{w}_k$: $\hat{w}_k = h(a)$ for $k = 1, \ldots, t-1$, $\hat{w}_k = \tilde{w}_k/c$ for $k = t, \ldots, s-1$, and $\hat{w}_k = h(b)$ for $k = s, \ldots, r$. By (A.7),

$$\sum_{k=1}^{t-1} p_k h(a) c \geq \sum_{k=1}^{t-1} p_k \tilde{w}_k = \sum_{k=1}^{t-1} z_k,$$

and hence

$$\begin{aligned} &\sum_{k=1}^{t-1} (z_k - c p_k \hat{w}_k)(\xi_k - \hat{\xi}_k) = \sum_{k=1}^{t-1} (z_k - p_k h(a) c)(\xi_k - a) \\ \leq &\sum_{k=1}^{t-1} (z_k - p_k \tilde{w}_k)(\xi_k - a) = \sum_{k=1}^{t-1} (z_k - p_k \tilde{w}_k)\xi_k \leq 0, \end{aligned}$$

where the first inequality follows from $\xi_k \geq a$ and $\tilde{w}_k < h(a)c$ for $k = 1, \ldots, t-1$; the second equality follows from (A.7). The second inequality follows from (A.6) with ξ in (A.6) replaced by $(\xi_1 + \Psi(\tilde{w}_1), \ldots, \xi_{t-1} + \Psi(\tilde{w}_{t-1}), \Psi(\tilde{w}_t), \ldots, \Psi(\tilde{w}_{r-r_0}))$, where Ψ is any nondecreasing function for which $\xi_{t-1} + \Psi(\tilde{w}_{t-1}) \leq \Psi(\tilde{w}_t)$. Next,

$$\sum_{k=t}^{s-1} (z_k - c p_k \hat{w}_k)(\xi_k - \hat{\xi}_k) = \sum_{k=t}^{s-1} (z_k - p_k \tilde{w}_k)(\xi_k - h^{-1}(\tilde{w}_k/c)) \leq 0,$$

where the inequality follows from (A.6) with ξ in (A.6) replaced by $(\Psi(\tilde{w}_1), \ldots, \Psi(\tilde{w}_{t-1}), \xi_t, \ldots, \xi_{s-1}, \Psi(\tilde{w}_s), \ldots, \Psi(\tilde{w}_{r-r_0}))$ and with $\Psi(x) = h^{-1}(x/c)$. For the last segment, note that

$$\begin{aligned} &\sum_{k=s}^{r-r_0} (z_k - c p_k \hat{w}_k)(\xi_k - \hat{\xi}_k) = \sum_{k=s}^{r-r_0} (z_k - p_k h(b) c)(\xi_k - b) \\ = &(\xi_{r-r_0} - b) \sum_{k=s}^{r-r_0} (z_k - p_k h(b) c) - \sum_{j=s+1}^{r-r_0} (\xi_j - \xi_{j-1}) \sum_{k=s}^{j-1} (z_k - p_k h(b) c), \end{aligned}$$

by summation by parts. On the other hand,

$$h(b)c < \tilde{w}_s = \max_{i \leq s} \min_{s \leq j \leq r-r_0} \frac{z_i + \cdots + z_j}{p_i + \cdots + p_j} = \min_{s \leq j \leq r-r_0} \frac{z_i + \cdots + z_j}{p_i + \cdots + p_j},$$

since x_s is a jump point of $\tilde{w}$. So,

$$\sum_{k=s}^{t} z_k \geq h(b) c \sum_{k=s}^{t} p_k, \quad \text{for all } \ t = s, \ldots, r - r_0.$$

That

$$\sum_{k=s}^{r-r_0} (z_k - c p_k \hat{w}_k)(\xi_k - \hat{\xi}_k) \leq 0, \quad \text{for all } \xi \in \Xi_{a,b},$$

follows immediately, and (A.4) is an easy consequence from summing the three segments.

Case (iii). When there are only two of the three groups of $\hat{w}_k$, (A.4) can be checked similarly as in Case (ii). The theorem follows.

Applications to l and l^*_α. Let $z_k = n_k/n$, $h(x) = e^x$ and $\beta(w) = \sum_{i=0}^r w_i p_i$. When N is known, let $g = l$, $r_0 = 0$, $a = -\infty$ and $c = H(w) = (N-n)/[n(1-\beta(w))]$. When N is unknown, take $g = l^*_\alpha$, $r_0 = 1$, $\epsilon = h(a) \geq \alpha > 0$ and $c = H(w) = [\beta(w) - \alpha]/\beta^2(w)$. Then $\hat{w}$ in (5) and (11) are the (penalized) maximum likelihood estimator of w given θ, by Theorem 3.

References

Bickel, P. J., Nair, V. N. and Wang, P. C. C. (1992). Nonparametric inference under biased sampling from a finite population. *Ann. Statist.* **20**, 853-878.

Bloomfield, P., Deffeyes, K. S., Silverman, B., Watson, G. S., Benjamini, Y. and Stine, R. A. (1979). Volume and area of oil fields and their impact on order of discovery. *Resource Estimation and Validation Project, Department of Statistics and Geology, Princeton University.*

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research.* Lyon : International Agency for Research on Cancer.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory.* Hayward, Calif. Institute of Mathematical Statistics.

Cook, R. D. and Martin, F. B. (1974). A model for quadrat sampling with "visibility bias". *J. Amer. Statist. Assoc.* **69**, 345-349.

Csörgő, M. and Revesz, P. (1981). *Strong Approximation in Probability and Statistics.* Academic Press.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069-1112.

Groeneboom, P. (1985). Estimating a monotone density. In *Proc. Conf. in Honor of Jerzy Neyman and Jack Kiefer* **2** (Edited by L. M. LeCam and R. A. Olshen), 539-555.

Groeneboom, P. and Pyke, R. (1983). Asymptotic normality of statistics based on the convex minorants of empirical distribution functions. *Ann. Probab.* **11**, 328-345

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255-277.

Gordon, L. (1993). Estimation for large successive samples with unknown inclusion probabilities. *Adv. Appl. Math.* **14**, 89-122.

Lynden-Bell, D. (1993). Eddington-Malmquist bias, streaming motions, and the distribution of galaxies. In *Statistical Challenges in Modern Astronomy* (Edited by G. J. Babu and E. D. Feigelson), 201-220.

Manski, C. F. (1993). The selection problem in econometrics and statistics. In *Econometrics* (Edited by G. S. Maddala, C. R. Rao and H. D. Vinod), 73-84. Amsterdam; North-Holland, New York.

Meisner, J. and Demirmen F. (1981). The creaming method: a Bayesian procedure to forecast future oil and gas discoveries in mature exploration provinces. *J. Roy. Statist. Soc. Ser. A.* **144**, 1-31.

Patil, G. P. and Rao, C. R. (1976). The weighted distributions: a survey of their applications. In *Applications of Statistics. Proceedings of the symposium held at Wright State University,*

Dayton, Ohio, 14-18 *June* 1976 (Edited by P. R. Krishnaiah), 383-405. North-Holland, Amsterdam.

Robertson, T., Wright, F. and Dykstra R. (1988). *Order Restricted Inference*. John Wiley.

Robbins, H. and Zhang, C. H. (1988). Estimating a treatment effect under biased sampling. *Proc. Nat. Acad. Sci.* **85**, 3670-3672.

Silverman, B. W. (1986). *Density Estimation*. Chapman and Hall.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616-620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13**, 178-205.

Vardi, Y. and Zhang, C. H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *Ann. Statist.* **20**, 1022-1039.

Woodroofe, M. (1993). Discussion of "Eddington Malmquist bias, streaming motions, and the distribution of galaxies". In *Statistical Challenges in Modern Astronomy* (Edited by G. J. Babu and E. D. Feigelson), 217-220.

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

Sun, J. and Woodroofe, M. (1996). Adaptive smoothing for a penalized npmle of a non-increasing density. *J. Statist. Plann. Inference* **52**, 143-159.

Wu, C. F. J. (1983). Convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

Department of Statistics, Case Western Reserve University, Cleveland, OH 44106, U.S.A.
E-mail: jiayang@sun.cwru.edu

E-mail: michaelw@stat.lsa.umich.edu

(Received November 1994; accepted February 1996)