A Hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator From Censored Data

Jon A. Wellner; Yihui Zhan

Journal of the American Statistical Association, Vol. 92, No. 439 (Sep., 1997), 945-959.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199709%2992%3A439%3C945%3AAHAFCO%3E2.0.CO%3B2-S

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/astata.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Thu Mar 10 06:17:37 2005

A Hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator From Censored Data

Jon A. WELLNER and Yihui ZHAN

We present a hybrid algorithm for nonparametric maximum likelihood estimation from censored data when the log-likelihood is concave. The hybrid algorithm uses a composite algorithmic mapping combining the expectation-maximization (EM) algorithm and the (modified) iterative convex minorant (ICM) algorithm. Global convergence of the hybrid algorithm is proven; the iterates generated by the hybrid algorithm are shown to converge to the nonparametric maximum likelihood estimator (NPMLE) unambiguously. Numerical simulations demonstrate that the hybrid algorithm converges more rapidly than either of the EM or the naive ICM algorithm for doubly censored data. The speed of the hybrid algorithm makes it possible to accompany the NPMLE with bootstrap confidence bands.

KEY WORDS: Algorithm; Censoring; EM algorithm; Hybrid method; Iterative convex minorant; Missing data; Self-consistency.

1. INTRODUCTION

This article presents a new general approach to iterative computation of the nonparametric maximum likelihood estimator (NPMLE) of a distribution function F on R when the observations can be viewed as incomplete data with concave log-likelihood. Because the iterations of our new algorithm are generated by a composite algorithmic mapping alternating steps of a (modified) iterative convex minorant (ICM) algorithm and of an expectation-maximization (EM) algorithm, we call it a *hybrid algorithm.* The hybrid algorithm is remarkable partly because it is simple and possesses global convergence and partly because it taps into the different strengths of both the EM algorithm and the modified ICM algorithm.

If the EM algorithm converges, then limit points of the iterates generated by the algorithm will satisfy the self-consistency equations. As the information loss due to censoring or missing data becomes heavier, the self-consistency equations tend to have multiple solutions. For example, whereas the self-consistency equations have a unique solution in the case of right-censored data, this is no longer the case for doubly censored data; the self-consistency equations may have more than one solution in the case of double censoring. A self-consistent estimate is not necessarily the NPMLE. (See Gu and Zhang 1993, p. 612, for an example of this with doubly censored data.) As another example, the self-consistency equations for interval-censored data do not determine uniquely the NPMLE either (see Groeneboom and Wellner 1992). In these cases the self-consistency equations fail to characterize the NPMLE.

A first consequence of this fact is that the EM algorithm becomes ambiguous in the sense that it may converge to a solution of the self-consistency equations other than the NPMLE. In fact, for certain observed data patterns, initial conditions lead the EM algorithm to a self-consistent estimate that is not the NPMLE, and for the same observed data patterns, other initial conditions lead the EM algorithm to the NPMLE. Mykland and Ren (1995) coped with this problem in the case of double-censoring by characterizing the NPMLE and altering the initial conditions of the EM algorithm to calculate a particular self-consistent estimate satisfying the characterizing conditions.

A second consequence is that heavy information loss results in a slow convergence rate of the EM algorithm. The arguments of Meilijson (1989) showed that the convergence rate of the EM algorithm in a missing-data problem depends on the ratio between incomplete data information and the complete data information. From Meilijson (1989, formula 12, p. 132), it follows that

$$\theta^{(m)} - \hat{\theta} \approx (\mathbf{I} - \mathbf{I}_X^{-1}\mathbf{I}_Y)^m(\theta^{(0)} - \hat{\theta}),$$

where $\mathbf{I}_X$ and $\mathbf{I}_Y$ denote the information matrices for θ in the complete data and incomplete data problems and $\mathbf{I}$ is the identity matrix of the same dimension. Although the EM algorithm performs satisfactorily in the case of right-censored (or even doubly censored) data (in which case $(\mathbf{I} - \mathbf{I}_X^{-1}\mathbf{I}_Y) < \mathbf{I}$), there is a general empirical finding that the number of iterations for the EM algorithm to compute the NPMLE for interval-censored data (in which case $\mathbf{I} - \mathbf{I}_X^{-1}\mathbf{I}_Y = \mathbf{I}$) increases with sample size (Groeneboom and Wellner 1992). Even though the EM algorithm may be fast enough to calculate the NPMLE itself (once), it may be far too slow to implement bootstrap methods (which involve computation of the NPMLE *many times* from resampled data).

Several different strategies to speed up the EM algorithm have been proposed, including Aitken acceleration, quasi-Newton methods, and conjugate gradient methods. (See Jamshidian and Jennrich 1993 for an approach using conjugate gradient acceleration and a nice review.)

Jon A. Wellner is Professor, Department of Statistics, University of Washington, Seattle, WA 98195. Yihui Zhan is Research Scientist, Mathsoft Statsci Division, Seattle, WA 98109. This work has been partially supported by National Science Foundation grant DMS-9306809, NIAID grant 2R01 AI291968-04, and NATO NWO grant B61-238. The authors thank Piet Groeneboom for introducing them to iterative convex minorant algorithms, and Geurt Jongbloed for explaining his results and pointing out the appropriate convergence results for composite algorithms.

© 1997 American Statistical Association

Our approach is somewhat different, instead involving characterizations of the NPMLE via Fenchel duality and the related convex minorants—characterizations that have their origin in the work of Ayer, Brunk, Ewing, Reid, and Silverman (1955) and Van Eeden (1956) in connection with "interval-censoring" models. As an alternative to the EM algorithm in these models with heavy information loss, Groeneboom (1991) proposed the ICM (see Groeneboom and Wellner 1992, pp. 65–74). This method is based on the simple observation that a distribution function F must be nondecreasing together with the fact that the log-likelihood function is concave in many missing-data problems. The NPMLE thus can be explicitly characterized via Fenchel duality for convex optimization. In many cases the characterization can be expressed as the left derivative of the convex minorant for a cumulative sum diagram defined by the derivative processes of the likelihood functions (Groeneboom 1996; Groeneboom and Wellner 1992; Huang and Wellner 1995a,b; Zhan and Wellner 1995). This is because the characterization can be equivalently viewed as the isotonic regression of the derivative processes under a monotonicity constraint, which possesses a convex minorant interpretation.

Jongbloed (1995a, 1995b) proposed a *modified ICM algorithm* with a line search and proved that the modified algorithm is globally convergent. The ICM algorithm is particularly suitable for computation of the NPMLE with interval-censored data and other problems in which the large-sample theory of the estimator involves normalization not by the square root of the sample size, but instead by the cube root of the sample size or another slower rate, and the EM algorithm often exhibits a slow convergence rate (as an algorithm for fixed sample size).

Motivated by the need to calculate the NPMLE unambiguously and efficiently in models with different information loss, and by the need to calculate the NPMLE quickly to implement bootstrap methods, we present a new hybrid algorithm consisting of alternating steps of the ICM and EM algorithms. The ICM step is based on the characterization of the NPMLE that identifies a particular solution to the self-consistency equations that maximizes the likelihood function. Heuristically, the hybrid algorithm is designed to search for the NPMLE, by the ICM scheme, among the set of all self-consistent estimates specified by the EM iterations. Because the set of all self-consistent estimates is a small subset of all feasible estimates, the hybrid algorithm should be able to outperform the naive ICM algorithm or the EM algorithm, independent of the information loss in a given problem.

An advantage of using a composite mapping of the ICM and the EM algorithm is that the algorithmic mapping of the EM iteration never destroys the ascent likelihood function in the modified ICM algorithm. This allows us to establish the global convergence of the hybrid algorithm by using a general convergence theorem for composite algorithmic mappings, theorem 7.3.4 of Bazaraa, Sherali, and Shetti (1993).

To demonstrate the effectiveness of the hybrid algorithm, we carried out detailed simulation experiments on the double-censoring model. The reason we chose the double-censoring model for simulation is the representativeness of this model; it reduces to both the right-censoring model and the interval-censoring model (case 1, or current status data) by appropriate choice of the censoring mechanism.

2. THE EXPECTATION-MAXIMIZATION ALGORITHM AND SELF-CONSISTENCY

The type of model that we treat here is as follows. Suppose that X is a real-valued random variable with distribution function F on R. Unfortunately, we are not able to observed X itself, but instead observe only $\mathbf{Y} = \mathbf{T}(X, \mathbf{C})$, where $\mathbf{C}$ is a random vector in $\mathbb{R}^m$ that is independent of X and $\mathbf{T}$ is a (measurable) function from $\mathbb{R} \times \mathbb{R}^m$ to $\mathbb{R}^k$ for some k.

We are interested in the NPMLE of F based on observation of a sample of $\mathbf{Y}$'s, $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ iid all with the same distribution as $\mathbf{Y}$.

Example 2.1: Right Censoring. Suppose that X and C take values in $\mathbb{R}^+ \equiv [0, \infty)$ and let $\mathbf{Y} = \mathbf{T}(X, C) \equiv (X \wedge C, 1_{[X \le C]}) \in \mathbb{R}^+ \times \{0, 1\}$.

Example 2.2: Interval Censoring, Case 1 (Current Status Data). Suppose that X and C take values in $\mathbb{R}^+$ and that $\mathbf{Y} = \mathbf{T}(X, C) \equiv (C, 1_{[X \le C]}) \in \mathbb{R}^+ \times \{0, 1\}$.

Example 2.3: Double Censoring. Suppose that X takes values in $\mathbb{R}^+$, $\mathbf{C}$ takes values in $R^{+2}_{\le} \equiv \{(u, v) \in \mathbb{R}^2 : 0 \le u \le v < \infty\}$. Then in this model we observe

$$\mathbf{Y} = \mathbf{T}(X, \mathbf{C}) \equiv \left\{ \begin{array}{ll} (X, 1) & \text{if } C_1 < X \le C_2 \\ (C_2, 2) & \text{if } X > C_2 \\ (C_1, 3) & \text{if } X \le C_1 \end{array} \right\} \equiv (W, \Delta).$$

Example 2.4: Interval Censoring, Case 2. Suppose that X takes values in $\mathbb{R}^+$ and $\mathbf{C}$ takes values in $R^{+2}_{\le} \equiv \{(u, v) \in \mathbb{R}^2 : 0 \le u \le v < \infty\}$ as in Example 2.3, but now we can only observe

$$\mathbf{Y} = \mathbf{T}(X, \mathbf{C}) = (C_1, C_2, 1_{[0, C_1]}(X), 1_{(C_1, C_2]}(X), 1_{(C_2, \infty)}(X)) \equiv (\mathbf{C}, \boldsymbol{\Delta}).$$

Example 2.5: Multiplicative Censoring. Suppose that both X and C take values in $\mathbb{R}^+$ and C, with the distribution G of C being the mixture $p\delta_1 + (1 - p)G_0$, where G_0 is a fixed (continuous) distribution on $\mathbb{R}^+$ (and δ_1 denoting the distribution concentrated at 1). Suppose that we observe $\mathbf{Y} = \mathbf{T}(X, C) = (XC, 1_{[C=1]})$. When $p = 0$, this is pure "multiplicative censoring"—the distribution of Y_1 is a scale mixture of G_0 with mixing distribution F.

A common feature of the foregoing examples on which we want to focus is as follows. There is some finite set of points $\{W_{(j)}\}_{j=1}^s \subset \mathbb{R}$ with $W_{(1)} \le \cdots \le W_{(s)}$ depending only on the coordinates of $Y_1, \ldots, Y_n$, so it is reasonable to assume that the NPMLE F_n of F is a discrete (sub) distribution function with jumps only at the points $\{W_{(j)}\}$.

For example, in Examples 2.1–2.3, the $W_{(j)}$'s are the order statistics of the first component of the Y_i's, and $s = n$ (if there are no ties; otherwise, $s \leq n$); in Example 2.4, the $W_{(j)}$'s are obtained by deleting the obviously irrelevant values (for the likelihood) from the first two components of the Y_i's, the pairs (C_{1i}, C_{2i}), then forming the union of the resulting set of real values (of random size $s \leq 2n$), and ordering them. Thus $s \leq 2n$ is random in this case. (Note that this is equivalent to forming the $T_{(j)}$'s in definition 1.1 of Groeneboom and Wellner 1992, p. 45.) In Example 2.5, take the $W_{(j)}$'s to be the ordered values of the first component of the Y_i's.

Thus our convention throughout the remainder of this article is as follows: by the NPMLE of the distribution function F, we mean a discrete (possibly sub) distribution function that maximizes the likelihood over the set $\mathcal{D}$ of discrete distribution functions that are piecewise constant between the points $W_{(1)} \leq \cdots \leq W_{(s)}$. This is an assumption for the current setting, which restricts our methods to a (large) subclass of such problems. In some cases it is of interest to maximize the likelihood over *all* distributions, or perhaps (e.g., Example 2.5 with $p = 0$) over all discrete distributions with the locations of the jumps as well as the magnitudes of the jumps varying.

Let $F \in \mathcal{D}$ be a distribution function that is piecewise constant between the points $W_{(j)}$, $j = 1, 2, \ldots, s$. Any $F \in \mathcal{D}$ can have jumps only at those points. A function $F \in \mathcal{D}$ can be identified with a vector $\mathbf{p} = (p_1, \ldots, p_s, p_{s+1})^T$, where $p_j = F(W_{(j)}) - F(W_{(j-1)})$ is the jump of F at $W_{(j)}$ for $j = 1, 2, \ldots, s$ and p_{s+1} represents the possibly remaining mass. The feasible set for $\mathbf{p}$ is

$$\mathbf{C_p} = \left\{ \mathbf{p} \in [0,1]^{s+1} \colon \sum_{j=1}^{s+1} p_j = 1, \text{and } p_j \geq 0, j \geq 1 \right\} \subset \mathbb{R}^{s+1}. \quad (1)$$

A function $F \in \mathcal{D}$ can also be identified with a vector $\mathbf{x} = (x_1, \ldots, x_s)$, where $x_j = F(W_{(j)})$ is the value of F evaluated at $W_{(j)}$. Correspondingly, the feasible set for $\mathbf{x}$ is

$$\mathbf{C_x} = \{\mathbf{x} \colon 0 \leq x_1 \leq \cdots \leq x_s \leq 1\} \subset \mathbb{R}^s. \quad (2)$$

Obviously, there is a one-to-one correspondence between a vector $\mathbf{p}$ and a vector $\mathbf{x}$ parameterizing the same function $F \in \mathcal{D}$. This correspondence will be conveniently denoted by $\mathbf{x} = \mathbf{x}(\mathbf{p})$ or $\mathbf{p} = \mathbf{p}(\mathbf{x})$, referring to the corresponding parameterization in $\mathbf{x}$ resulting from $\mathbf{p}$ and that in $\mathbf{p}$ resulting from $\mathbf{x}$.

The EM algorithm was originally formulated in the context of parametric estimation problems, but there are many examples of its use in nonparametric settings (see, e.g., Efron 1967, Laird 1978, Tsai and Crowley 1985, Turnbull 1974, 1976, Vardi 1982, 1989, and Vardi and Zhang 1992). Although the NPMLE rarely has a solution in "closed form," the EM algorithm is usually straightforward to implement. Moreover, all of the limit points generated by the EM iterations can be characterized in terms of self-consistency equations, a set of facts that we now briefly review.

Let $\mathbf{p}^{(0)}$ be an initial estimate of F, and let $\mathbf{p}^{(m)}$ denote the current estimate of F after m steps of the algorithm. The log-likelihood for the complete data is

$$\begin{aligned} l(\mathbf{p}, X_1, \ldots, X_n) &= \log \prod_{i=1}^{n} f(X_i|\mathbf{p}) \\ &= \sum_{j=1}^{s+1} \#\{X_i = W_{(j)}\} \log p_j. \end{aligned}$$

In the E step of the EM algorithm, we compute the conditional expectation of the complete-data log-likelihood given the iid observations $\mathbf{Y}_i$, $i = 1, 2, \ldots, n$:

$$\begin{aligned} Q(\mathbf{p}|\mathbf{p}^{(m)}) &= E_{F^{(m)}}[l(\mathbf{p}, X_1, \ldots, X_n)|\mathbf{Y}_1, \ldots, \mathbf{Y}_n] \\ &= \sum_{j=1}^{s+1} (\log p_j) \sum_{i=1}^{n} P_{F^{(m)}}\{X_i = W_{(j)}|\mathbf{Y}_i\}. \quad (3) \end{aligned}$$

In the last expression, $P_{F^{(m)}}$ denotes the conditional probability distribution of X given Y when $X \sim F^{(m)}$.

The M step in the EM algorithm can be calculated explicitly in a missing-data problem. Maximize (3) over the feasible region $\mathbf{C_p}$ defined in (1). To do this, let $\alpha_{ij} = P_{F^{(m)}}\{X_i = W_{(j)}|Y_i\}$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, s + 1$. Write $\alpha_{\cdot j} = \sum_{i=1}^{n} \alpha_{ij}$ and $\alpha_{\cdot\cdot} = \sum_{j=1}^{s+1} \alpha_{\cdot j} = n$. The Lagrangian function for this problem can be written as

$$\begin{aligned} L(\mathbf{p}, \lambda) &= \sum_{j=1}^{s+1} (\log p_j) \sum_{i=1}^{n} \alpha_{ij} + \lambda \left(1 - \sum_{j=1}^{s+1} p_j\right) \\ &= \sum_{j=1}^{s+1} \alpha_{\cdot j} \log p_j + \lambda \left(1 - \sum_{j=1}^{s+1} p_j\right). \end{aligned}$$

Differentiating with respect to p_j and setting the derivatives to 0 yields $\alpha_{\cdot j}/p_j - \lambda = 0$, or $p_j = \alpha_{\cdot j}/\lambda$ for $j = 1, 2, \ldots, s + 1$. Because $\sum_{j=1}^{s+1} p_j = 1$ and $\alpha_{\cdot\cdot} = n$, we obtain $\lambda = \alpha_{\cdot\cdot} = n$. This leads us to the optimum point $\mathbf{p}^{(m+1)}$ with its components

$$p_j^{(m+1)} = \frac{\alpha_{\cdot j}}{n} = \frac{1}{n} \sum_{i=1}^{n} P_{F^{(m)}}\{X_i = W_{(j)}|\mathbf{Y}_i\} \quad (4)$$

for $j = 1, 2, \ldots, s + 1$.

The EM iteration actually leads to a self-consistent estimate if the iteration converges. Rewriting Equation (4) in its cumulative form in terms of distribution functions yields

$$F^{(m+1)}(x) = E_{F^{(m)}}\left[\frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \leq x]}|\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right]$$

for all $x \in \mathbb{R}$. Thus it is clear that the limit, $F^{(\infty)}$ say, of the sequence of functions $\{F^{(m)}\}$ will satisfy

$$F^{(\infty)}(x) = E_{F^{(\infty)}}\left[\frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \leq x]}|\mathbf{Y}_1, \ldots, \mathbf{Y}_n\right], x \in \mathbb{R}. \quad (5)$$

This is Efron's (1967) property of *self-consistency of an estimator* $F^{(\infty)}$ of F. The equations in (5) are called the *self-consistency equations.* Thus if the EM algorithm converges, then it will converge to a solution of the self-consistency equations; that is, a self-consistent estimate.

The nonparametric likelihood equations in the context of missing-data problems also reduce to the self-consistency equations. For a bounded function h, we define a curve (or parametric submodel) $\{F_\eta\}$ passing through F by

$$\frac{dF_\eta}{dF}(X) = 1 + \eta\left(h(X) - \int h\, dF\right)$$

where $(h - \int h\, dF)$ is the centered score for the complete-data model because

$$\frac{\partial}{\partial\eta}\left(\log\frac{dF_\eta}{dF}(X)\right)\Big|_{\eta=0} = h(X) - \int h\, dF.$$

But the score function $B(F)(h)(Y)$ for F based on incomplete data is the conditional expectation of the centered score in the complete-data model [Bickel, Klaassen, Ritov, and Wellner 1993, prop. A.5.5 (A)]:

$$B(F)(h)(Y) = E_F\left[h(X) - \int h\, dF | Y\right].$$

Setting $h(x) \equiv h_t(x) = 1_{[0,t]}(x)$ following Gill (1989), we obtain the nonparametric score functions for iid incomplete observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$:

$$\begin{aligned}\Psi_n(F)(h_t) &\equiv \frac{1}{n}\sum_{i=1}^n B(F)(h_t)(\mathbf{Y}_i)\\ &= E_F\left[\frac{1}{n}\sum_{i=1}^n 1_{[0,t]}(X_i) - F(t) | \mathbf{Y}_1, \ldots, \mathbf{Y}_n\right]\\ &= E_F\left[\frac{1}{n}\sum_{i=1}^n 1_{[0,t]}(X_i) | \mathbf{Y}_1, \ldots, \mathbf{Y}_n\right] - F(t).\end{aligned}$$

Thus the nonparametric likelihood equations $\Psi_n(F)(h_t) = 0$, $t \in R$, are exactly the self-consistency equations. The likelihood equations do not provide any additional information to distinguish between different self-consistent estimates.

Example 2.3, (cont.): Double Censoring. An explicit EM iteration for a self-consistent estimate in the double-censoring model is available. We calculate

$$\begin{aligned}\alpha_{ij} &= P_{F^{(m)}}\{X_i = W_{(j)} | (W_i, \Delta_i)\} = 1_{[\Delta_i=1]} \cdot 1_{[W_i = W_{(j)}]}\\ &\quad + 1_{[\Delta_i=2]}\frac{p_j^{(m)}}{1 - F^{(m)}(W_i)} 1_{[W_{(j)} > W_i]}\\ &\quad + 1_{[\Delta_i=3]}\frac{p_j^{(m)}}{F^{(m)}(W_i)} 1_{[W_{(j)} \le W_i]} \qquad (6)\end{aligned}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n+1$; here $W_{(n+1)}$ is a point located to the right of $W_{(n)}$ for the remaining mass.

Let $H_n^{(j)}(t) = n^{-1}\sum_{i=1}^n 1_{[W_i \le t, \Delta_i = j]}$ be the empirical subdistribution functions corresponding to $\Delta_i = j$, $j = 1, 2, 3$. Let $H_n(t) = n^{-1}\sum_{i=1}^n 1_{[W_i \le t]}$ be the empirical distribution function for the W_i's. It is also easy to show that a self-consistent estimate F_n of F for doubly censored data is specified by the equations

$$\begin{aligned}F_n(t) = H_n(t) &- \int_{[0,t]} \frac{1 - F_n(t)}{1 - F_n(u)}\, dH_n^{(2)}(u)\\ &+ \int_{(t,\infty)} \frac{F_n(t)}{F_n(u)}\, dH_n^{(3)}(u) \qquad (7)\end{aligned}$$

for all $t \in [0, \infty)$. Mykland and Ren (1996) used the discrete version of this equation to construct an equivalent fixed-point problem and thereby connect back to the EM algorithm.

3. THE NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATOR AND ITERATIVE CONVEX MINORANT ALGORITHMS

Our convention here is that the NPMLE of F is a discrete distribution function $F_n \in \mathcal{D}$ maximizing the log-likelihood

$$\phi(\mathbf{x}) = \log\left(\prod_{i=1}^n g(\mathbf{Y}_i | \mathbf{p}(\mathbf{x}))\right).$$

Formally, the NPMLE $\hat{\mathbf{x}}$ is defined by the following optimization problem:

$$\begin{cases}\max \log\left(\prod_{i=1}^n g(\mathbf{Y}_i | \mathbf{p}(\mathbf{x}))\right)\\ \text{over } \mathbf{x} \in \mathbf{C_x} = \{\mathbf{x} \in [0,1]^s : 0 \le x_1 \le \cdots \le x_s \le 1\}.\end{cases} \qquad (8)$$

Let $D_\mathbf{x} = \{\mathbf{x} : x_1 \le \cdots \le x_s\}$. We can extend the definition of ϕ to $D_\mathbf{x}$ by defining $\phi(\mathbf{x}) = -\infty$ for $\mathbf{x} \in D_\mathbf{x} \setminus \mathbf{C_x}$. Then the optimization problem (8) can be equivalently stated as

$$\begin{cases}\max \phi(\mathbf{x})\\ \text{over } \mathbf{x} \in D_\mathbf{x}.\end{cases} \qquad (9)$$

3.1 Characterization of the Nonparametric Maximum Likelihood Estimator

For the optimization problems stated in the previous section, necessary and sufficient conditions characterizing the NPMLE can be given explicitly by use of the Fenchel (or Lagrange) duality theorem. The following lemma is from Jongbloed (1995a,b) as reformulated slightly by Groeneboom (1996).

Lemma 3.1. Let ϕ: $\mathbb{R}^n \mapsto \mathbb{R} \cup \{-\infty\}$ be a continuous concave function. Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex cone, and let $\mathcal{K}_0 = \mathcal{K} \cap \phi^{-1}(\mathbb{R})$. Suppose that $\mathcal{K}_0$ is nonempty and ϕ is differentiable on $\mathcal{K}_0$. Then $\hat{\mathbf{x}} \in \mathcal{K}_0$ satisfies

$$\phi(\hat{\mathbf{x}}) = \max_{x \in \mathcal{K}} \phi(x) \qquad (10)$$

if and only if

$$\langle \hat{\mathbf{x}}, \nabla\phi(\hat{\mathbf{x}})\rangle = 0 \tag{11}$$

and

$$\langle \mathbf{x}, \nabla\phi(\hat{\mathbf{x}})\rangle \le 0 \tag{12}$$

for all $\mathbf{x} \in \mathcal{K}$.

Example 2.3 (cont.): Double Censoring. After eliminating trivial situations, the maximization problem defining the NPMLE can be equivalently stated as

$$\begin{cases} \phi(\mathbf{x}) = \sum_{i=1}^{s}\{1_{[\Delta_{(i)}=1]}\log(x_i - x_{i-1}) \\ \qquad\qquad + \delta_{(i)}\log(1-x_i) + 1_{[\Delta_{(i)}=3]}\log x_i\} \\ \text{over } \mathbf{x} \in \mathbf{C}_{\mathbf{x}} = \{\mathbf{x}: 0 \le x_1 \le \cdots \le x_s \le 1\}, \end{cases} \tag{13}$$

where the integer $s = n$ if $\Delta_{(n)} \neq 1$ and $s = n-1$ if $\Delta_{(n)} = 1$, and the numbers $\delta_{(i)}$ are defined by

$$\delta_{(i)} = \begin{cases} 1_{[\Delta_{(i)}=2]} & \text{if } i = 1, 2, \ldots, s-1 \\ 1_{[\Delta_{(n-1)}=2]} + 1_{[\Delta_{(n)}=1]} & \text{if } i = s. \end{cases}$$

Strictly speaking, the function $\phi(\mathbf{x})$ is not exactly the log-likelihood function in this model. But the difference between the log-likelihood function and $\phi(\mathbf{x})$ is a constant that does not depend on $\mathbf{x}$. To see this, note that the likelihood for doubly censored sample $\{(W_i, \Delta_i)\}_{i=1}^n$ is given by

$$C + \sum_{i=1}^{n}\{1_{[\Delta_{(i)}=1]}\log(x_i - x_{i-1}) + 1_{[\Delta_{(i)}=2]} \times \log(1-x_i) + 1_{[\Delta_{(i)}=3]}\log x_i\}, \tag{14}$$

where C is a constant not depending on $\mathbf{x}$.

In the likelihood (14), we may assume, without loss of generality, that $\Delta_{(n)} \neq 3$. Actually, if $\Delta_{(j)} = 3$ for $j = k, k+1, \ldots, n$, then we can take $x_j = 1$ for $k \le j \le n$ to make $1_{[\Delta_{(j)}=3]}\log x_j$ as large as possible (namely 0), without putting any additional constraint on other components of $\mathbf{x}$.

When $\Delta_{(n)} \neq 3$, the only possibility is that either $\Delta_{(n)} = 1$ or $\Delta_{(n)} = 2$. If $\Delta_{(n)} = 1$, then the maximizing $\mathbf{x}$ must have its last component $\hat{x}_n = 1$ to make the term $1_{[\Delta_{(n)}=1]}\log(\hat{x}_n - \hat{x}_{n-1})$ as large as possible without putting extra constraint on other components of $\mathbf{x}$. Then the term $1_{[\Delta_{(n)}=1]}\log(1-\hat{x}_{n-1})$ can be combined with the term $1_{[\Delta_{(n-1)}=2]}\log(1-\hat{x}_{n-1})$ to form the term $\delta_{(n)}\log(1-\hat{x}_{n-1})$.

In the maximization problem (14), we may also assume that $\Delta_{(1)} \neq 2$. The reason is that if $\Delta_{(j)} = 2$ for $j = 1, 2, \ldots, k$, then we can take $x_j = 0$ for $1 \le j \le k$ to make the corresponding term $1_{[\Delta_{(j)}=2]}\log(1-x_j)$ in ϕ as large as possible (namely 0) without putting any additional constraints on other components of $\mathbf{x}$.

It is worthwhile to note that the NPMLE without assuming that $\Delta_{(1)} \neq 2$ and $\Delta_{(n)} \neq 3$ is generally a function consisting of three parts. If $\Delta_{(1)} = \cdots = \Delta_{(m_0-1)} = 2$ and $\Delta_{(m_0)} \neq 0$, then $F_n(W_{(1)}) = \cdots = F_n(W_{(m_0-1)}) = 0$ and $F_n(W_{(m_0)}) > 0$. If $\Delta_{(m_1+1)} = \Delta_{(m_1+2)} = \cdots = \Delta_{(n)} = 3$, $\Delta_{(m_1)} \neq 3$, then $F_n(W_{(m_1+1)}) = \cdots = F_n(W_{(n)}) = 1$. The third part, corresponding to $\Delta_{(m_0)}, \ldots, \Delta_{(m_1)}$ is characterized by Lemma 3.1 as follows:

Theorem 3.1. Suppose that $\Delta_{(1)} \neq 2$ and $\Delta_{(n)} \neq 3$. Let $\Delta_{(s+1)} = 0$. Then $\hat{\mathbf{x}}$ maximizes ϕ over the feasible set $\mathbf{C}_{\mathbf{x}}$ defined in (13) iff

$$\frac{1_{[\Delta_{(k)}=1]}}{\hat{x}_k - \hat{x}_{k-1}} - \sum_{i=k}^{s}\frac{\delta_{(i)}}{1-\hat{x}_i} + \sum_{i=k}^{s}\frac{1_{[\Delta_{(i)}=3]}}{\hat{x}_i} \times \begin{cases} \le 0, & \forall k = 1, 2, \ldots, s \\ = 0, & \text{if } \hat{x}_k > \hat{x}_{k-1}. \end{cases} \tag{15}$$

Moreover, $\hat{\mathbf{x}}$ is the unique point that maximizes $\phi(\mathbf{x})$ over the feasible set $\mathbf{C}_{\mathbf{x}}$.

The proof of the theorem starts with expressing a point $\mathbf{x} \in \mathbf{C}_{\mathbf{x}}$ by

$$\mathbf{x} = \sum_{i=1}^{s}(x_i - x_{i-1})\mathbf{1}_i$$

with $x_0 = 0$, where the vector $\mathbf{1}_i$ has 1's as its last $s - i + 1$ components and 0's as its first $i-1$ components, $i = 1, 2, \ldots, s$. Note that

$$\sum_{j=k}^{s}\frac{\partial\phi(\hat{\mathbf{x}})}{\partial x_j} = \frac{1_{[\Delta_{(k)}=1]}}{\hat{x}_k - \hat{x}_{k-1}} - \sum_{i=k}^{s}\frac{\delta_{(i)}}{1-\hat{x}_i} + \sum_{i=k}^{s}\frac{1_{[\Delta_{(i)}=3]}}{\hat{x}_i}.$$

Then the condition (3.5) in Lemma 3.1 reduces to

$$\langle \mathbf{x}, \nabla\phi(\hat{\mathbf{x}})\rangle = \sum_{k=1}^{s}(x_k - x_{k-1})\sum_{j=k}^{s}\frac{\partial\phi(\hat{\mathbf{x}})}{\partial x_j} \le 0$$

for all $\mathbf{x} \in \mathbf{C}_{\mathbf{x}}$. This implies the first condition given in (15). Similarly, the condition (11) in Lemma 3.1 reduces to the second condition in (15).

The uniqueness of the NPMLE $\hat{\mathbf{x}}$ is derived from the fact that the Hessian matrix $\partial^2\phi(\mathbf{x})/\partial\mathbf{x}\partial\mathbf{x}^T$ of the objective function ϕ is negative definite (see the Appendix, Zhan 1996, or Zhan and Wellner 1995).

3.2 The Iterative Convex Minorant Algorithm

The idea behind the (ICM) algorithm can be seen from the following equivalent relationship.

Theorem 3.2. Suppose that ϕ satisfies the condition in Lemma 3.1. For any positive definite matrix $\boldsymbol{\Sigma}$, a point $\mathbf{x}^* \in \mathcal{K}$ maximizes the quadratic function

$$J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}}) \equiv -\frac{1}{2}[\mathbf{x} - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}\nabla\phi(\hat{\mathbf{x}})]^T \times \boldsymbol{\Sigma}[\mathbf{x} - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}\nabla\phi(\hat{\mathbf{x}})]$$

over the convex cone $\mathcal{K}$ iff $\mathbf{x}^* = \hat{\mathbf{x}} \equiv \arg\max_{\mathbf{x}\in\mathcal{K}}\phi(\mathbf{x})$.

Maximizing $J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}})$ over $\mathcal{K}$ is "easier" than maximizing ϕ over $\mathcal{K}$, assuming that we know $\hat{\mathbf{x}}$. Because we do not know $\hat{\mathbf{x}}$, this is in fact a bit unrealistic, but it still gives

a valuable perspective. To continue this line of thought, suppose that we choose $\boldsymbol{\Sigma}$ to be a positive definite diagonal matrix; then maximizing $J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}})$ over $\mathcal{K}$ possesses a convex minorant interpretation. Furthermore, if ϕ is twice continuously differentiable and $\boldsymbol{\Sigma} = -\boldsymbol{\Sigma}(\hat{\mathbf{x}})$ is chosen to be the Hessian matrix of ϕ at $\hat{\mathbf{x}}$, maximizing $J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}})$ can also be interpreted as maximizing a quadratic approximation $\tilde{\phi}$ of ϕ at the point $\hat{\mathbf{x}}$:

$$\begin{aligned}\tilde{\phi}(\mathbf{x}) &= \phi(\hat{\mathbf{x}}) + [\mathbf{x} - \hat{\mathbf{x}}]^T \nabla\phi(\hat{\mathbf{x}}) \\ &\quad + \frac{1}{2}[\mathbf{x} - \hat{\mathbf{x}}]^T[-\boldsymbol{\Sigma}(\hat{\mathbf{x}})][\mathbf{x} - \hat{\mathbf{x}}] \\ &= \text{const.} - \frac{1}{2}[\mathbf{x} - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}(\hat{\mathbf{x}})\nabla\phi(\hat{\mathbf{x}})]^T \\ &\quad \times \boldsymbol{\Sigma}(\hat{\mathbf{x}})[\mathbf{x} - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}(\hat{\mathbf{x}})\nabla\phi(\hat{\mathbf{x}})].\end{aligned}$$

This interpretation motivates the following iterative scheme. Let $\Omega\{\hat{\mathbf{x}}\}$ be the solution set, set $k \leftarrow 0$:

1. If $\mathbf{x}^{(k)} \in \Omega$ then stop.
2. Choose $\mathbf{x}^{(k+1)}$ to be any point in a set $A(\mathbf{x}^{(k)})$ such that $J_{\boldsymbol{\Sigma}(\mathbf{x}^{(k)})}(\mathbf{x}, \mathbf{x}^{(k)})$ is maximized; $k \leftarrow k+1$ and go to 1.

The set $A(\mathbf{x}^{(k)})$ can be viewed as the image of a point-to-set mapping A that maps a point $\mathbf{x}^{(k)}$ to a set $A(\mathbf{x}^{(k)})$ of all maximizers of $J_{\boldsymbol{\Sigma}(\mathbf{x}^{(k)})}(\mathbf{x}, \mathbf{x}^{(k)})$. The mapping A, called the *algorithmic mapping,* is formally defined by

$$A(\mathbf{x}) = \arg\max_{\mathbf{z} \in \mathcal{K}} J_{\boldsymbol{\Sigma}(\mathbf{x})}(\mathbf{z}, \mathbf{x}).$$

With ϕ a continuously differentiable function on the set $\{\mathbf{x} \in \mathcal{K}\colon \phi(\mathbf{x}) > -\infty\}$, it is trivially true that the algorithmic mapping $A(\mathbf{x})$ is well defined on the same set $\{\mathbf{x} \in \mathcal{K}\colon \phi(\mathbf{x}) > -\infty\}$. Moreover, $A(\mathbf{x})$ is continuous at each point $\mathbf{x} \in \mathcal{K}$ where $\phi(\mathbf{x}) > -\infty$ and $\mathbf{x} \mapsto \boldsymbol{\Sigma}(\mathbf{x})$ is continuous. Although an ascent function does not exist in general, the direction generated by the ICM algorithm at any point $\mathbf{x} \neq \hat{\mathbf{x}}$ is a direction of ascent of the objective function ϕ. The following lemma due to Jongbloed (1995a,b) states this property.

Lemma 3.2. Let $\phi\colon \mathbb{R}^n \mapsto [-\infty, \infty)$ be a function satisfying the condition of Lemma 3.1. Then for all λ sufficiently small,

$$\phi(\mathbf{x} + \lambda(A(\mathbf{x}) - \mathbf{x})) > \phi(\mathbf{x})$$

for any $\mathbf{x} \in \mathcal{K} - \{\hat{\mathbf{x}}\}$ and any positive definite matrix $\boldsymbol{\Sigma}(\mathbf{x})$.

Based on this lemma, we may introduce a line search between the iterations in the ICM algorithm to produce a feasible point and an ascent objective function, thereby guaranteeing global convergence of the algorithm. The algorithm with a particular line search determined by a variant of Armijio's rule is called the *modified ICM algorithm* and is due to Jongbloed (1995a,b).

4. A HYBRID ALGORITHM

As a starting point, we review the algorithmic mapping of the modified ICM algorithm.

Let the segment $\text{seg}(\mathbf{x}, \mathbf{z})$ be defined as $\{\mathbf{w}\colon \mathbf{w} = \mathbf{x} + \lambda(\mathbf{z} - \mathbf{x}), 0 \leq \lambda \leq 1\}$. For fixed $0 < \varepsilon < 1/2$, let the mapping $N(\mathbf{x})$ be defined by

$$N(\mathbf{x}) = \begin{cases} \{A(\mathbf{x})\} & \text{if } A(\mathbf{x}) \text{ satisfies condition in (17)} \\ \mathbf{z} \in \mathbf{C}(\mathbf{x}) & \text{elsewhere.} \end{cases} \tag{16}$$

In the last expression, the first condition is given by

$$\phi(A(\mathbf{x})) > \phi(\mathbf{x}) + (1 - \varepsilon)\nabla\phi(\mathbf{x})^T(A(\mathbf{x}) - \mathbf{x}), \tag{17}$$

and the set $\mathbf{C}(\mathbf{x})$ is defined by

$$\begin{aligned}\mathbf{C}(\mathbf{x}) = \{\mathbf{z} \in \text{seg}(\mathbf{x}, A(\mathbf{x})) &: (1 - \varepsilon)\nabla\phi(\mathbf{x})^T(\mathbf{z} - \mathbf{x}) \\ &\geq \phi(\mathbf{z}) - \phi(\mathbf{x}) \geq \varepsilon\nabla\phi(\mathbf{x})^T(\mathbf{z} - \mathbf{x})\}.\end{aligned}$$

The new *hybrid algorithm* is generated iteratively by a composition mapping $M \cdot N$ of the EM algorithm and the modified ICM algorithm: Set $k \leftarrow 0$:

1. If $\mathbf{x}^{(k)} \in \Omega$ then stop.
2. $\mathbf{x}^{(k+1)} \in MN(\mathbf{x}^{(k)})$; $k \leftarrow k+1$ and go to 1.

In this iteration, the algorithmic mapping $M(\cdot)$ is defined by the EM iteration

$$Q(M(\mathbf{p})|\mathbf{p}) \geq Q(\mathbf{p}'|\mathbf{p}) \tag{18}$$

for every vector $\mathbf{p}' \in \mathbf{C}_\mathbf{p}$.

A very nice property of the EM algorithm is that it generates a sequence of ascent log-likelihood functions in the iteration. A formal description of this property involves some related quantities. Let

$$k(x_1, \ldots, x_n|y_1, \ldots, y_n, \mathbf{p}) = \frac{f(x_1, \ldots, x_n|\mathbf{p})}{g(y_1, \ldots, y_n|\mathbf{p})} = \prod_{i=1}^{n} \frac{f(x_i|\mathbf{p})}{g(y_i|\mathbf{p})}$$

denote the conditional density of the complete data $X_1, \ldots, X_n$ given the incomplete data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ and $\mathbf{p}$. In these expressions, the incomplete-data specification $g(y|\mathbf{p})$ and the complete-data specification $f(x|\mathbf{p})$ are related by

$$g(\mathbf{y}|\mathbf{p}) = \int_{\mathbf{y}=T(x,c)} f(x|\mathbf{p})\, dx\, dG(c).$$

The log-likelihood $\phi(\mathbf{x}(\mathbf{q})) \equiv L(\mathbf{q})$ parameterized in $\mathbf{q}$ can be rewritten as

$$L(\mathbf{q}) = \log \prod_{i=1}^{n} g(\mathbf{Y}_i|\mathbf{q}) = Q(\mathbf{q}|\mathbf{p}) - H(\mathbf{q}|\mathbf{p}),$$

where the functions Q and H are defined by

$$Q(\mathbf{q}|\mathbf{p}) = E_\mathbf{p}[\log f(X_1, \ldots, X_n|\mathbf{q})|\mathbf{Y}_1, \ldots, \mathbf{Y}_n]$$

and

$$H(\mathbf{q}|\mathbf{p}) = E_\mathbf{p}[\log k(X_1, \ldots, X_n|\mathbf{Y}_1, \ldots, \mathbf{Y}_n, \mathbf{q})|\mathbf{Y}_1, \ldots, \mathbf{Y}_n],$$

which are assumed to exist for all pairs of $(\mathbf{p}, \mathbf{q})$.

Because $H(\mathbf{p}|\mathbf{p}) \geq H(\mathbf{q}|\mathbf{p})$ for any $\mathbf{q}$ in (1) (see lemma 1 in Dempster, Laird, and Rubin 1977), it follows from (18) that for any $\mathbf{p} \in \mathbf{C}_\mathbf{p}$,

$$\begin{aligned}L(M(\mathbf{p})) &= Q(M(\mathbf{p})|\mathbf{p}) - H(M(\mathbf{p})|\mathbf{p}) \\ &\geq Q(\mathbf{p}|\mathbf{p}) - H(\mathbf{p}|\mathbf{p}) = L(\mathbf{p}).\end{aligned}$$

Let $\mathbf{y} = \mathbf{x}(\mathbf{q})$ and $\mathbf{x} = \mathbf{x}(\mathbf{p})$ denote the parameterization of $\mathbf{q}$ and $\mathbf{p}$ in the feasible set $\mathbf{C}_\mathbf{x}$. The monotonicity $L(M(\mathbf{p})) \geq L(\mathbf{p})$ at any point $\mathbf{p} \in \mathbf{C}_\mathbf{p}$ implies that

$$\phi(\mathbf{y}) \geq \phi(\mathbf{x}) \tag{19}$$

for any $\mathbf{y} \in M(\mathbf{x})$ and any $\mathbf{x} \in \mathbf{C}_\mathbf{x}$. For convenience of notation, we have used the corresponding parameterization $\mathbf{y}$ of $\mathbf{q}$ and the corresponding parameterization $\mathbf{x}$ of $\mathbf{p}$ in the EM algorithmic mapping $M(\cdot)$.

4.1 Global Convergence of the Hybrid Algorithm

There is a general theorem on the global convergence of an algorithm generated by a composite algorithmic mapping. Basically, if one of the algorithmic mappings yields an ascent objective function and is closed, while the other does not destroy the ascent function, then the algorithm generated by the composite mapping of the two mappings converges. We now give a statement of this result, theorem 7.3.4 of Bazaraa et al. (1993). (The proof can be found in Bazaraa et al. 1993, p. 253.)

Theorem 4.1. Let $\mathbf{X}$ be a nonempty closed set in $\mathbb{R}^n$ and let $\Omega \subset \mathbf{X}$ be a nonempty solution set. Let α: $\mathbb{R}^n \mapsto \mathbb{R}$ be a continuous function and consider the point-to-set mapping $\mathbf{C}$: $\mathbf{X} \mapsto \mathbf{X}$ satisfying the following property: Given $\mathbf{x} \in \mathbf{X}$, it holds that $\alpha(\mathbf{y}) \leq \alpha(\mathbf{x})$ for $\mathbf{y} \in \mathbf{C}(\mathbf{x})$. Let $\mathbf{B}$: $\mathbf{X} \mapsto \mathbf{X}$ be a point-to-set mapping that is closed over the complement of Ω and that satisfies $\alpha(\mathbf{y}) < \alpha(\mathbf{x})$ for each $\mathbf{y} \in \mathbf{B}(\mathbf{x})$, if $\mathbf{x} \notin \Omega$. Now consider the algorithm defined by the composite mapping $\mathbf{A} = \mathbf{CB}$. Given $\mathbf{x}^{(1)} \in \mathbf{X}$, suppose that the sequence $\{\mathbf{x}^{(k)}\}$ is generated as follows:

1. If $\mathbf{x}^{(k)} \in \Omega$ then stop.
2. Let $\mathbf{x}^{(k+1)} \in \mathbf{A}(\mathbf{x}^{(k)})$; $k \leftarrow k+1$ and go to 1.

Suppose that the set $\Lambda = \{\mathbf{x}\colon \alpha(\mathbf{x}) \leq \alpha(\mathbf{x}^{(1)})\}$ is compact. Then either the algorithm stops in a finite number of steps with a point in Ω, or all accumulation points of $\{\mathbf{x}^{(k)}\}$ belong to Ω.

To apply this theorem, we need to prove the closedness of the algorithmic mapping N as defined in the modified ICM algorithm and the existence of an ascent function. This is given in the following lemma.

Lemma 4.1. Suppose that the function ϕ is continuously differentiable on the set $\{\mathbf{x} \in \mathbb{R}^n\colon \phi(\mathbf{x}) > -\infty\}$. Let $\mathbf{x}^{(0)} \in \mathcal{K}$ satisfy $\phi(\mathbf{x}^{(0)}) > -\infty$ and let

$$K = \{\mathbf{x} \in \mathcal{K}\colon \phi(\mathbf{x}) \geq \phi(\mathbf{x}^{(0)})\}. \tag{20}$$

Assume that the mapping $\mathbf{x} \mapsto \mathbf{\Sigma}(\mathbf{x})$ is continuous in the sense that all elements of $\mathbf{\Sigma}(\mathbf{x})$ are continuous on the set K. Then

1. The algorithmic mapping $N(\cdot)$ defined in (16) is closed at each point $\mathbf{x} \in K - \{\hat{\mathbf{x}}\}$.
2. For all $\mathbf{x} \neq \hat{\mathbf{x}}$ and for all $\mathbf{z} \in N(\mathbf{x})$, it holds that $\phi(\mathbf{z}) < \phi(\mathbf{x})$.

The global convergence of the hybrid algorithm is shown in the following theorem by an application of Theorem 4.1.

Theorem 4.2. Suppose that the function ϕ: $\mathbb{R}^n \mapsto [-\infty, \infty)$ satisfies the following conditions:

1. ϕ is concave and attains its maximum over $\mathcal{K}$ at a unique point $\hat{\mathbf{x}}$.
2. ϕ is continuously differentiable on the set $\{\mathbf{x} \in \mathbb{R}^n\colon \phi(\mathbf{x}) > -\infty\}$.

Suppose that $\mathbf{x}^{(0)} \in \mathcal{K}$ satisfies $\phi(\mathbf{x}^{(0)}) > -\infty$, and that the mapping $\mathbf{x} \mapsto \mathbf{\Sigma}(\mathbf{x})$ is continuous on the set $\{\mathbf{x} \in \mathcal{K}\colon \phi(\mathbf{x}) \geq \phi(\mathbf{x}^{(0)})\}$. Then the hybrid algorithm generated by the composite mapping of the modified ICM and EM algorithms converges to the NPMLE $\hat{\mathbf{x}}$.

A few simple properties of the NPMLE $\hat{\mathbf{x}}$ are worth recording.

Proposition 4.1. If the NPMLE $\hat{\mathbf{x}}$ is unique in the sense that $\hat{\mathbf{x}}$ is a unique point in $\mathbf{C}_\mathbf{x}$ that maximizes ϕ: $\phi(\hat{\mathbf{x}}) > \phi(\mathbf{x})$ for any other $\mathbf{x} \neq \hat{\mathbf{x}}$ in the feasible set $\mathbf{C}_\mathbf{x}$, then the NPMLE $\hat{\mathbf{x}}$ is a fixed point of the EM algorithmic mapping $M(\cdot)$, the modified ICM algorithmic mapping $N(\cdot)$, and the hybrid algorithmic mapping $MN(\cdot)$.

Corollary 4.1. Under the condition of this proposition, the NPMLE $\hat{\mathbf{x}}$ satisfies the self-consistency equation.

4.2 The Hybrid Algorithm for Doubly Censored Data

Maximization of $J_{\mathbf{\Sigma}(\mathbf{x})}(\mathbf{w}, \mathbf{x})$ with a positive definite diagonal matrix $\mathbf{\Sigma}(\mathbf{x})$ has a convex minorant interpretation. Let $\mathbf{\Sigma}(\mathbf{x}) = \operatorname{diag}\{d_1, \ldots, d_s\}$ be a diagonal matrix with positive constants $d_i > 0$ possibly depending on $\mathbf{x}$, $i = 1, 2, \ldots, s$. Let $r_i = x_i + \nabla\phi_i/d_i$, $i = 1, 2, \ldots, s$. Then

$$J_{\mathbf{\Sigma}(\mathbf{x})}(\mathbf{w}, \mathbf{x}) = -\frac{1}{2}\sum_{i=1}^{s} d_i(w_i - r_i)^2.$$

The characterization condition for $\hat{\mathbf{x}}$ to maximize $J_{\mathbf{\Sigma}(\mathbf{x})}(\mathbf{w}, \mathbf{x})$ over $\mathbf{C}_\mathbf{x}^+$ is given by Lemma 3.1 as

$$\sum_{i=k}^{s}(\hat{x}_i - r_i)\begin{cases} \geq 0, & \text{for } k = 1, 2, \ldots, s \\ = 0, & \text{if } \hat{x}_k > \hat{x}_{k-1}. \end{cases}$$

The cumulative sum diagram is defined by the following points:

$$P_i = \left(\sum_{j=1}^{i} d_j, \sum_{j=1}^{i} d_j r_j\right), \qquad i = 1, 2, \ldots, s.$$

The convex minorant is defined as the greatest convex piecewise linear function lying below the points P_i, $i = 0, 1, 2, \ldots, s$ with $P_0 \equiv (0, 0)$.

Let $0 = i_0 < i_1 < \cdots < i_p = s$ be the indices on the x-axis corresponding to the vertices of the convex minorant. Note that the slopes $\hat{x}_i$ of the convex minorant remain constant between the integer block $(i_{k-1}, i_k]$: $\hat{x}_i \equiv \hat{x}_{i_k}$, $i = i_{k-1} + 1, \ldots, i_k$ for $k = 1, 2, \ldots, p-1$. This, together with the fact that these slopes are equal to the slopes of the cumulative sum diagram on the corresponding block

$(i_{k-1}, i_k]$, imply that

$$\sum_{i=i_k}^{s} d_i(\hat{x}_i - r_i) = 0, \qquad k = 1, 2, \ldots, p-1.$$

This is exactly the second part of the characterization condition.

Similarly, the fact that these slopes $\hat{x}_i$ are greater than or equal to the slopes of the cumulative sum diagram on the block $(j, s]$ implies that

$$\sum_{i=j}^{s} d_i(\hat{x}_i - r_i) \geq 0, \qquad j = 1, 2, \ldots, s.$$

This is exactly the first part of the characterization condition. The foregoing argument follows the lines of theorem 1.5 of Barlow, Bartholemew, Bremner, and Brunk (1972) (see also Groeneboom and Wellner 1992, Jongbloed 1995a, and Robertson, Wright, and Dykstra 1988).

Because the log-likelihood function in the double-censoring model is actually twice differentiable on the set $\{\mathbf{x}\colon \phi(\mathbf{x}) > -\infty\}$, we may choose $\boldsymbol{\Sigma}(\mathbf{x})$ to be a diagonal matrix consisting of the diagonal elements in the negative Hessian matrix of ϕ and calculate the NPMLE by the ICM algorithm. Unlike the interval censoring case 1, where the negative Hessian matrix of the objective function is of a diagonal form so that the ICM actually maximizes a quadratic approximation to the objective function, in the case of double-censoring the ICM only maximizes a linear approximation to ϕ with a diagonal quadratic approximation.

Example 2.3 (cont.): Double Censoring. For any vector $\mathbf{x} = (x_1, \ldots, x_s)^T$ in the feasible set $\mathbf{C}_\mathbf{x}$ defined in (3.6), define the processes $G(\mathbf{x}, \cdot)$ and $V(\mathbf{x}, \cdot)$ by $G(\mathbf{x}, 0) = 0$,

$$\begin{aligned} G(\mathbf{x}, j) &= \sum_{i=1}^{j} (-\nabla^2 \phi_{ii}(x)) \\ &= \sum_{i=1}^{j} \left[\frac{1_{[\Delta_{(i)}=1]}}{(x_i - x_{i-1})^2} + \frac{1_{[\Delta_{(i+1)}=1]}}{(x_{i+1} - x_i)^2} \right. \\ &\qquad \left. + \frac{\delta_{(i)}}{(1 - x_i)^2} + \frac{1_{[\Delta_{(i)}=3]}}{x_i^2} \right] \end{aligned} \tag{21}$$

for $j = 1, 2, \ldots, s$; and $V(\mathbf{x}, 0) = 0$,

$$V(\mathbf{x}, j) = \sum_{i=1}^{j} [x_i(-\nabla^2 \phi_{ii}(x)) + \nabla \phi_i(x)] \tag{22}$$

for $j = 1, 2, \ldots, s$. In the foregoing, $\Delta_{(s+1)} = 0$ and $\nabla^2 \phi_{ij}(\mathbf{x}) = \partial^2 \phi(\mathbf{x}) / \partial x_i \partial x_j$. Let $\boldsymbol{\Sigma}(\mathbf{x}) = -[\nabla^2 \phi_{ij}(\mathbf{x})]$ be the diagonal part of the negative of the Hessian of ϕ.

Theorem 4.3. A point $\mathbf{x}^*$ is the NPMLE or, equivalently, maximizes $J_{\boldsymbol{\Sigma}(\hat{\mathbf{x}})}(\mathbf{w}, \hat{\mathbf{x}})$ over $\mathbf{C}_\mathbf{x}^+$, iff it is the left derivative of the convex minorant of the cumulative sum diagram consisting the following points:

$$P_0 = (0, 0), \qquad P_j = (G(\hat{\mathbf{x}}, j), V(\hat{\mathbf{x}}, j)), \qquad j = 1, 2, \ldots, m.$$

Theorem 4.3 is a consequence of Theorem 3.2. It is the motivation behind the ICM algorithm for doubly censored data. Although it gives only a first-order approximation, it is simple to implement. The EM iteration for doubly censored data has a closed form; see (4). Putting these together, we now give the iteration steps of the hybrid algorithm.

Let $\Delta_{i,l}(\mathbf{x}) = (V(\mathbf{x}, i) - V(\mathbf{x}, l))/(G(\mathbf{x}, i) - G(\mathbf{x}, l))$. The iteration steps of the hybrid algorithm are as follows:

Step 1. Choose an initial guess $\mathbf{x}^{(0)}$. One example might be $\mathbf{x}^{(0)}$ with all components equal to $1/(s+1)$.

Step 2. For each iterate $\mathbf{x}^{(k)}$, $k \geq 0$, compute the weight process $G(\mathbf{x}^{(k)}, \cdot)$ and $V(\mathbf{x}^{(k)}, \cdot)$ and form the cumulative sum diagram and its convex minorant. Then compute the left derivative $\bar{\mathbf{x}}^{(k+1)}$ of the convex minorant of the cumulative sum diagram:

Step 2.1. Set $i_0 = 0$ and construct the set of indices

$$i_0 < i_1 < \cdots < i_l = s$$

such that

$$\Delta_{i_j, i_{j-1}}(\mathbf{x}^{(k)}) = \min\{\Delta_{r, i_{j-1}}(\mathbf{x}^{(k)}) : i_{j-1} < r \leq s\}$$

for $j = 1, \ldots, l = l(k)$.

Step 2.2. Set

$$x_i^{(k+1)} = \Delta_{i_j, i_{j-1}}(\mathbf{x}^{(k)})$$

for $i = i_{j-1} + 1, \ldots, i_j$.

Step 3. For a fixed $0 < \varepsilon < 1/2$, if $\bar{\mathbf{x}}^{(k)}$ satisfies

$$\phi(\bar{\mathbf{x}}^{(k)}) > \phi(\mathbf{x}^{(k)}) + (1 - \varepsilon) \nabla \phi(\mathbf{x}^{(k)})^T (\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}),$$

then let $\hat{\mathbf{x}}^{(k)} = \bar{\mathbf{x}}^{(k)}$ and go to Step 4; otherwise, perform a line search on the segment $\text{seg}(\mathbf{x}^{(k)}, \bar{\mathbf{x}}^{(k)})$, and obtain $\hat{\mathbf{x}}^{(k)}$ such that

$$\phi(\hat{\mathbf{x}}^{(k)}) - \phi(\mathbf{x}^{(k)}) \geq \varepsilon \nabla \phi(\mathbf{x}^{(k)})^T (\hat{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)})$$

and

$$\phi(\hat{\mathbf{x}}^{(k)}) - \phi(\mathbf{x}^{(k)}) \leq (1 - \varepsilon) \nabla \phi(\mathbf{x}^{(k)})^T (\hat{\mathbf{x}}^{(k)} - \mathbf{x}^{(k)}).$$

Step 4. One EM step (composition or hybridization step). Compute

$$p_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} P_{\hat{F}^{(k)}}\{X_i = W_{(j)} | (W_i, \Delta_i)\}, \qquad j = 1, 2, \ldots, n+1,$$

where $\hat{F}^{(k)}$ is the distribution function corresponding to $\hat{\mathbf{x}}^{(k)}$. Let $\mathbf{x}^{(k+1)}$ be the distribution function with its jumps given by $p_j^{(k+1)} = x_j^{(k+1)} - x_{j-1}^{(k+1)}$.

Step 5. Convergence testing. For the given convergence criterion $\|\cdot\|$ and the given tolerance ε, compute $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| = D^{(k+1)}$. If $D^{(k+1)} < \varepsilon$, then stop, and $\mathbf{x}^{(k+1)}$ is the NPMLE; otherwise, go to Step 1, with $\mathbf{x}^{(k+1)}$ the starting point.

In practice, the hybrid algorithm can be used without the line search needed in the modified ICM algorithm. Although no convergence of the algorithm is guaranteed, the simulation experiments that we performed in this article showed no problem of convergence. For numerical evidence in this direction, see the next section.

5. COMPARISON OF ALGORITHMS

There are at least three different types of convergence criteria. The first type is based on the distance between $F^{(k+1)}$ and $F^{(k)}$, such as the Euclidean distance or the maximum coordinatewise distance between the two m-dimensional vectors $x^{(k+1)}$ and $x^{(k)}$:

$$\|F_n^{(k+1)} - F_n^{(k)}\| = \max_{1\le i\le n} |x_i^{(k+1)} - x_i^{(k)}| < \varepsilon. \quad (23)$$

The second type is based on the likelihood ratio of $F^{(k+1)}$ and $F^{(k)}$, or, put another way, as the distance between the log-likelihoods $\phi(x^{(k+1)})$ and $\phi(x^{(k)})$:

$$|\phi(F_n^{(k+1)}) - \phi(F_n^{(k)})| < \varepsilon. \quad (24)$$

A third type of convergence criteria is based on the Fenchel conditions given in the characterization Theorem 3.1: a solution $\hat{x}$ is accepted as the NPMLE if

$$|\langle \nabla\phi(\hat{x}), \hat{x}\rangle| < \varepsilon$$

and

$$\max\left\{\sum_{i=k}^{s} \nabla\phi_i(\hat{x}) : k = 1, 2, \ldots, s\right\} < \varepsilon. \quad (25)$$

Because the Fenchel conditions characterize the NPMLE, this is perhaps the preferred convergence criterion. Our experience with the three classes of convergence criteria is that stopping iterations based on the likelihood ratio usually yields fewer iterations (although it may require some additional computation for the log-likelihood function evaluation), and stopping according to the Fenchel criterion usually requires more iterations than either of the other two criteria for the same ε. (For more examples, see Zhan and Wellner 1995.)

The following example illustrates that the EM algorithm might converge to a solution other than the NPMLE.

Example 5.1. Suppose that the observations are $\{(W, \Delta)\}_{i=1}^4 = \{(1,1),(2,2),(3,3),(4,3)\}$. It is easy to verify that the discrete distribution function assigning mass $2/3$ at 1 and $1/3$ at 4 is a self-consistent estimate but is not the NPMLE, which puts mass $1/2$ at 1 and $1/2$ at 3. In fact, the likelihood at the NPMLE is $-2\log(2) = -\log 4$, whereas the likelihood at the self-consistent estimate is $\log(4/27) \approx -\log 7$. The EM algorithm will converge to the self-consistent estimate $(\frac{2}{3}, \frac{2}{3}, \frac{2}{3}, 1)$ starting from the initial guess $x^{(0)} = (.1, .1, .1, .2)$. Initial conditions of the pattern (c, c, c, d) with $0 < c < d \le 1$ all resulted in the same self-consistent estimate, whereas starting the EM algorithm with $x^{(0)} = (.1, .1, .15, .2)$ led to the NPMLE $(\frac{1}{2}, \frac{1}{2}, 1, 1)$. Initial conditions of the pattern (c, c, b, d) with $0 < c < b < d$ all led to the NPMLE. On the other hand, the hybrid algorithm and the ICM algorithm starting with both types of initial conditions always converge to the NPMLE.

Table 1. Mean and Standard Deviations for Number of Iterations (Fenchel)

	Empirical	Kaplan—Meier	Interval case 1	Double censoring
ICM	1 ± 0	17.1 ± 9.3	3.6 ± 1.7	20.4 ± 15.5
Hybrid	1 ± 0	2.9 ± 2.2	2.0 ± .9	3.0 ± 2.4
EM	1 ± 0	17.9 ± 14.0	34.7 ± 30.5	20.4 ± 24.9

Example 5.2. The double-censoring model can be reduced to the right-censoring model, the interval-censoring model case 1, and the model involving no censoring at all. The performances of the ICM, the hybrid, and the EM algorithm are different from each other on these different models.

We started with a doubly censored sample of length $n = 5$:

$$\{(W_1, \Delta_1), \ldots, (W_5, \Delta_5)\} = \{(1,1),(2,1),(3,2),(4,2),(5,3)\}.$$

Because the number of iterations that the algorithms require to compute the NPMLE depends on the order of the 1s, 2s, and the 3s in the sample, we tested the algorithms on all 30 data configurations of the form

$$\{(W_1, \Delta_{\pi(1)}), \ldots, (W_5, \Delta_{\pi(5)})\}$$

for some permutation $\pi = (\pi(1), \ldots, \pi(5))$ of the integers $\{1, 2, 3, 4, 5\}$. Note that the original $\boldsymbol{\Delta}$ vector contains two 1s and two 2s, the total number of different configurations is $5!/(2!\,2!) = 30$.

The number of iterations for each algorithm to compute the NPMLE on each of the 30 sample configurations was then recorded. The means and standard deviations for the number of iterations over all 30 configurations are listed under the category "double censoring."

The sample configurations for the right-censoring model were constructed from the doubly censored samples by replacing the 3 with a 2. Again, all of the $C_2^5 = 10$ distinct sample configurations were tested. The means and standard deviations for the number of iterations are listed under "Kaplan–Meier." The interval-censored samples that are "comparable" to those in the right- and double-censored models were constructed by replacing one of the $\Delta = 1$'s in the doubly censored samples with a 2 and the other with a 3. Thus there are again $C_2^5 = 10$ distinct sample configurations to test. The means and standard deviations for the number of iterations are listed under the category "interval case 1."

For comparison, we also list the "means" and "standard deviations" of the number of iterations for the algorithms on a sample involving no censoring at all:

$$\{(W_1, \Delta_1), \ldots, (W_5, \Delta_5)\} = \{(1,1),(2,1),(3,1),(4,1),(5,1)\}.$$

Because the NPMLE in this case is the empirical distribution, the performance are listed under the category "empirical."

The performance of the three algorithms is illustrated in Table 1. The initial estimator for the three algorithms is the same: $(1/6, 2/6, \ldots, 5/6)^T$. The convergence criterion for the three algorithms are also the same: the Fenchel criterion defined in (25) with $\varepsilon = .0000001$. Of course, there are completely noniterative forms of the estimators in the first two columns (the empirical distribution function and

Table 2. Number of Iterations and User Times (Fenchel Criterion, $r \approx .6$)

Sample size n	500	1,000	1,500	2,000	2,500	3,000	4,000	5,000
EM	1,300	3,089	4,106	4,790	6,749	7,999	10,140	12,620
Hybrid	33	40	58	88	88	114	104	129
EM	55.33	583.92	1,780	3,651	7,982	13,529	37,191	60,461
Hybrid	2.42	8.84	24.46	56.0	89.92	157.58	280.10	513.88

the Kaplan–Meier estimator), so these comparisons are for illustration only.

It is clear from Table 1 that the hybrid algorithm needs fewer iterations to converge to the NPMLE than the other censoring models. It is significantly faster than the EM algorithm for all of the censoring models for sample size $n = 5$. The ICM also seems to perform more consistently (i.e., with smaller variability) than the EM algorithm.

In comparison to the ICM algorithm, the hybrid ICM-EM algorithm works significantly better for all the models involved regardless of the convergence criterion used for comparison. In particular, the hybrid ICM-EM algorithm performs slightly better than the ICM algorithm for the interval censoring case 1 model.

To illustrate the practical utility of the hybrid algorithm for larger sample sizes, we generated doubly censored data of length n (with n ranging from 500 to 5,000) from the exponential distribution with mean $\frac{1}{2}$: $F(x) = 1 - \exp(-2x)$. The censoring variables (C_1, C_2) were iid as $(U_{(k)}, U_{(l)})$ for $1 \le k < l \le 20$. The random variables $U_{(k)}$ and $U_{(l)}$ are the kth and the lth order statistics from a sample of $m = 20$ uniform $U(0, 1)$ random variables. We use the random number generator in S-PLUS 3.3 and the functions rexp( ) and runif( ) to obtain the random samples of X and (C_1, C_2). The EM algorithm, the ICM algorithm, and the ICM-EM hybrid algorithm are programmed in FORTRAN-77 with S-PLUS interfaces. The computations are carried out on a Sun SPARCStation 20.

Moderate Censoring. The censoring configuration is $k = 5$ and $l = 16$ in this case. This corresponds to a censoring rate $r \approx .60$ (the ratio between the number of censored observations and the sample size). The ICM-EM hybrid algorithm and the EM algorithm were applied to the same doubly censored samples, and the number of iterations needed to obtain convergence is recorded in Table 2. The user time (in seconds) of evaluating the S-PLUS functions for these algorithms are also listed in Table 2 (in the last two rows of the table). In programming the S-PLUS function, we tried to make sure that the preparations for both the hybrid algorithm and the EM algorithm are the same. For example, the samples are first reduced according to the discussions before Theorem 3.1. We implemented the algorithm on a Sun SPARCStation 20 using the S-PLUS 3.3 interface. The convergence criterion used for both the EM and the ICM-EM hybrid algorithm is the Fenchel convergence criterion defined in (25). The tolerance level is $\varepsilon = .0000001$. All the algorithms are started from the same initial estimator $x^{(0)} = (1/(s+1), 2/(s+1), \ldots, s/(s+1))^T$. As for the ICM, at the sample of $n = 500$ it took $510,974$ iterations to satisfy the Fenchel criterion. We did not carry the ICM through in the experiment.

We compare the EM algorithm equipped with the convergence criterion defined in (24) to the ICM-EM hybrid algorithm equipped with the same likelihood criterion defined in (24). The tolerance is $\varepsilon = .0000001$. Both the EM and the ICM-EM hybrid algorithm were started from the initial estimator $x^{(0)} = (1/(s+1), 2/(s+1), \ldots, s/(s+1))^T$. The number of iterations and the user times for the two algorithms to compute the NPMLE are shown in Table 3.

Both the number of iterations and the user times needed by the hybrid ICM-EM algorithm are again significantly smaller than that needed by the EM algorithm. The number of iterations for the naive ICM equipped with (24) as convergence criterion and started from the uniform initial estimate took too long to record and was omitted in the experiment. For example, for the sample of $n = 500$ in this comparison, 15,468 iterations were required to obtain the NPMLE for the tolerance of $\varepsilon = .000001$.

Heavy Censoring. To demonstrate the efficiency of the hybrid algorithm for highly censored samples, we take the censoring variables (C_1, C_2) to be iid as $(U_{(k)}, U_{(l)})$ for $k = 8$ and $l = 12$. The censoring rate is about .86 for this configuration. Samples of the same sizes were generated, and the EM and the hybrid algorithms were tested on each of them. The number of iterations and the user times were recorded in Table 4 (for EM algorithm). The convergence criterion used is the Fenchel criterion defined in (25), and the tolerance is again $\varepsilon = .0000001$.

Again, on the same sample, both the numbers of the iterations and the user times for the hybrid algorithm are significantly smaller than those of the EM algorithm. The saved number of iterations and the user time seem to increase with the sample size, indicating better performance

Table 3. Number of Iterations and User Times (Likelihood Criterion, $r \approx .6$)

Sample size n	500	1,000	1,500	2,000	2,500	3,000	4,000	5,000
EM	832	1,852	2,338	2,801	3,611	4,175	5,297	6,237
Hybrid	30	45	52	53	63	80	94	94
EM	36.48	356.05	1,012.81	2,154.12	4,334.79	7,184.76	17,363	29,614
Hybrid	2.33	9.65	23.48	40.26	72.47	124.48	261.05	410.34

Table 4. Number of Iterations and User Times (Fenchel Criterion, $r \approx .86$)

Sample size n	*500*	*1,000*	*1,500*	*2,000*	*2,500*	*3,000*	*4,000*	*5,000*
EM	2,090	3,927	5,889	7,344	8,507	5,473	14,390	14,161
Hybrid	45	79	106	120	125	143	205	124
EM	106.28	907.52	3,172	6,803	12,279	11,145	55,001	81,481
Hybrid	2.50	9.29	23.49	42.70	71.19	119.35	274.63	312.68

of the hybrid algorithm over the EM on large sample sizes.

Example 5.4. We generated doubly censored data from the same exponential distribution with mean $1/2$ and with the same censoring configuration $k = 5$ and $l = 16$. We generated a doubly censored sample of length $3,000$. There are $1,760$ observations with Δ being a 2 or a 3. The censoring rate is $1{,}760/3{,}000 = .59$ in this sample. The NPMLE together with the true distribution function is shown in Figure 1.

6. BOOTSTRAP CONFIDENCE BANDS FOR DOUBLY CENSORED DATA

Our motivation for the new hybrid algorithm was the ability to use the bootstrap with doubly censored data. Experience in using either the EM or ICM algorithm was that they were both too slow to enable implementing the bootstrap with moderate sample sizes. In this section we report a selection from simulation experiments that we have performed to illustrate the feasibility of bootstrap confidence band to accompany the NPMLE in the double-censoring model. Note that use of the bootstrap requires us to recompute the estimator many times based on bootstrap samples; hence speed of convergence of the computational algorithm becomes crucial. In this particular censoring model (double censoring), we expect that bootstrap confidence bands will "work well" (i.e., behave correctly asymptotically for large n) because of our knowledge of information bounds (see Bickel et al. 1993, sec. 6.6, ex. 6.6.5) and because of a proof of asymptotic validity of the bootstrap (see Wellner and Zhan 1996). It should be emphasized that we do not expect the following bootstrap methods to work well for all censoring problems, and in particular they are not likely to work well for problems involving interval censoring.

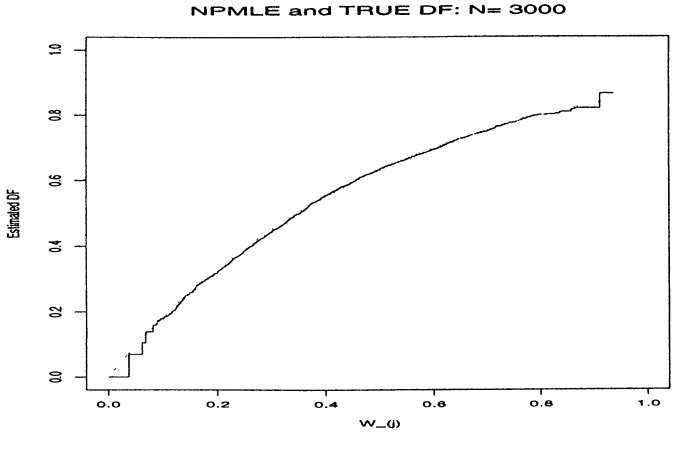

Figure 1. The NPMLE and True Exponential. The solid line is the NPMLE; the dotted line is the true exponential F.

Example 6.1: Bootstrap Confidence Bands and Achieved Levels To construct confidence bands for F based on doubly censored data using the bootstrap, we proceed as follows. We first compute the NPMLE F_n of F. Once F_n is available, we can also estimate the function $K = G_Y - G_Z$; because

$$H_P^{(2)}(t) = P\{W \le t, \Delta = 2\} = \int_{[0,t]} (1 - F(u))\, dG_Z(u)$$

and

$$H_P^{(3)}(t) = P\{W \le t, \Delta = 3\} = \int_{[0,t]} F(u)\, dG_Y(u),$$

the function $K(t)$ can also be written as

$$K(t) = G_Y(t) - G_Z(t) = \int_{[0,t]} \frac{dH_P^{(3)}(u)}{F(u)} - \int_{[0,t]} \frac{dH_P^{(2)}(u)}{1 - F(u)}.$$

Plugging in the empirical versions $H_n^{(j)}$ of $H_P^{(j)}$ for $j = 2, 3$ yields a natural estimate for the function K:

$$\hat{K}_n(t) = \int_{[0,t]} \frac{dH_n^{(3)}(u)}{F_n(u)} - \int_{[0,t]} \frac{dH_n^{(2)}(u)}{1 - F_n(u)}, \qquad (26)$$

where F_n is the NPMLE calculated from the same sample.

Now we are ready to bootstrap. For each of B bootstrap samples

$$(\hat{W}_{j1}, \hat{\Delta}_{j1}), \ldots, (\hat{W}_{jn}, \hat{\Delta}_{jn}), \qquad j = 1, \ldots, B,$$

from the empirical distribution of the observed data $(W_1, \Delta_1), \ldots, (W_n, \Delta_n)$, we compute $\hat{F}_{nj}$, and then

$$d_{nj} = \sqrt{n} \sup\{|\hat{K}_n(t)(\hat{F}_{nj}(t) - F_n(t))| : 0 \le t \le 1\}, \qquad j = 1, \ldots, B,$$

where the function $\hat{K}_n(t)$ is defined by (26). Then we approximate the critical value c_n of the bootstrap band by the $(1 - \alpha) \times 100\%$ percentile of these B numbers. The resulting bootstrap band is then given by

$$F_n(t) - \frac{c_n}{\sqrt{n}\hat{K}_n(t)} \le F(t) \le F_n(t) + \frac{c_n}{\sqrt{n}\hat{K}_n(t)}, \qquad 0 \le t \le W_{(n)}.$$

For illustration, consider survival times X generated from the exponential distribution with mean $1/2$, $F(x) = 1 - \exp(-2x)$, and censoring variables (C_1, C_2) (again) the 5th and the 16th order statistics from a uniform (0, 1) distribution. A 95% bootstrap confidence band for the exponential distribution in this example with sample size $n = 500$ and the number of bootstrap replications being $B = 500$ is

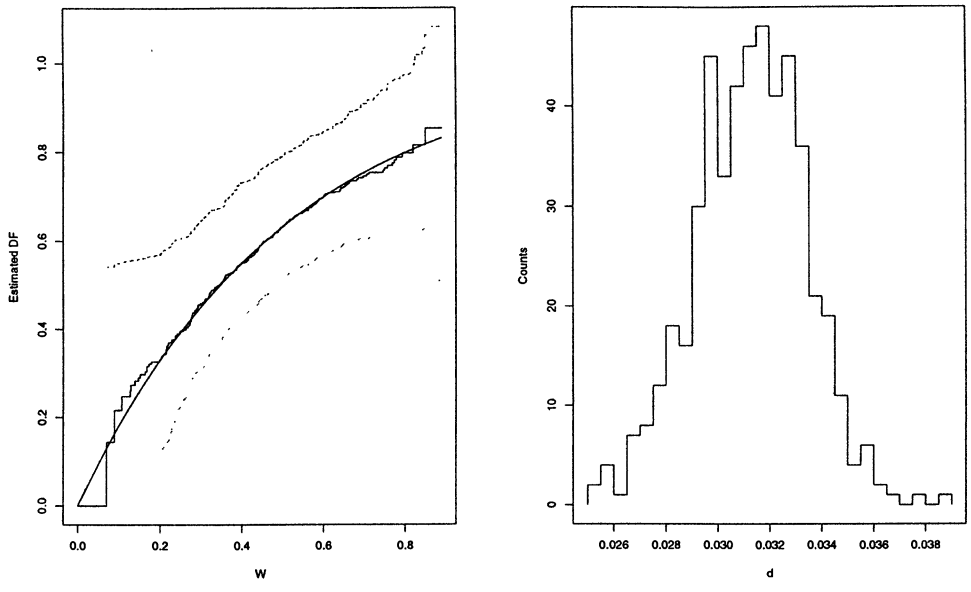

Figure 2. The Bootstrap Band for Exponential. (a) Bootstrap band: n = 500, alpha =.05; (b) histogram.

shown in Figure 2a. The histogram of the $B = 500$ distances d_n is shown in Figure 2b.

To get some feel for the achieved confidence level of the bootstrap confidence bands for moderate sample sizes, we carried out an experiment for sample size $n = 100$ using the exponential distribution with mean $1/2$ and censoring variables (C_1, C_2). As earlier, we generated doubly censored samples of size $n = 100$. A bootstrap confidence band with a given confidence level $(1 - \alpha)$ was then constructed based on this sample with the bootstrap sample size equal to the sample size n and the number of bootstrap replications being $B = 800$. Then we checked whether the true distribution function lies in the band so constructed. The number of simulations is 1,000. The proportion of the successful coverage among the 1,000 simulations is an estimate for the true confidence level of the bootstrap bands. Table 5 lists the achieved confidence levels for the experiment.

Apparently, the achieved confidence level is not too far from the nominal level, with a slight indication of conservativeness of the bootstrap bands in this case. Asymptotic theory justifying the bootstrap confidence bands for large sample sizes was given by Wellner and Zhan (1996).

7. DISCUSSION

We have proposed a new hybrid algorithm for computation of the NPMLE of the distribution function F of a real-valued random variable based on censored (missing) data. The new algorithm is based on alernating steps of the EM algorithm and of the ICM algorithm. Using results for optimization theory, we have shown that the new hybrid algorithm is globally convergent whenever started from a point with finite likelihood.

Table 5. Achieved Confidence Levels

α	.01	.05	.10	.15	.20
$\hat{\alpha}$	.013	.048	.097	.138	.183

Numerical exploration of the new hybrid algorithm in the case of double censoring shows that it converges to the NPMLE very quickly, beating both the EM and ICM algorithms in terms of number of iterations and computation time required. One way to view this heuristically is as follows. The ICM step of the algorithm searches for the NPMLE in the set of all self-consistent estimates specified by the EM iterations. Because the subset of all self-consistent estimates is a small subset of all feasible estimates, the hybrid algorithm improves on the naive ICM. The new hybrid algorithm makes feasible the implementation of bootstrap confidence bands to accompany the NPMLE.

APPENDIX: PROOFS

Proof of Lemma 3.1

We follow Jongbloed (1995a). Suppose that $\hat{\mathbf{x}} \in \mathcal{K}_0$ satisfies condition (12) and (11). Let $\mathbf{x} \in \mathcal{K}$ be arbitrary. Then by the concavity of ϕ we have

$$\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}}) \leq (\mathbf{x} - \hat{\mathbf{x}})^T \nabla\phi(\hat{\mathbf{x}}) \leq 0,$$

which implies that $\hat{\mathbf{x}}$ satisfies (10). Note that the inequality holds trivially if $x \in \mathcal{K}\backslash\mathcal{K}_0$.

Conversely, suppose that (10) holds. If condition (12) is not satisfied, then there is an $\mathbf{x} \in \mathcal{K}$ such that $\langle \mathbf{x}, \nabla\phi(\hat{\mathbf{x}})\rangle > 0$. Because $\mathcal{K}$ is a convex cone, it holds that

$$\hat{\mathbf{x}} + \varepsilon\mathbf{x} = (1+\varepsilon)\left(\frac{1}{1+\varepsilon}\hat{\mathbf{x}} + \left(1 - \frac{1}{1+\varepsilon}\right)\mathbf{x}\right) \in \mathcal{K}$$

for all $\varepsilon \geq 0$. It then follows (from continuity of ϕ and differentiability of ϕ on $\mathcal{K}_0$) that for $\varepsilon \downarrow 0$,

$$\phi(\hat{\mathbf{x}} + \varepsilon\mathbf{x}) - \phi(\hat{\mathbf{x}}) = \varepsilon\mathbf{x}^T\nabla\phi(\hat{\mathbf{x}}) + o(\varepsilon).$$

Hence we have $\phi(\hat{\mathbf{x}} + \varepsilon\mathbf{x}) > \phi(\hat{\mathbf{x}})$ when ε is sufficiently small, contradicting the assumption that $\hat{\mathbf{x}}$ maximizes ϕ over $\mathcal{K}$.

Now suppose that condition (11) is not satisfied; then $\hat{\mathbf{x}} \neq 0$ is true. For $|\varepsilon|$ sufficiently small, it holds that $(1+\varepsilon)\hat{\mathbf{x}} \in \mathcal{K}$. Taking the sign of ε the same as that of $\langle \hat{\mathbf{x}}, \nabla\phi(\hat{\mathbf{x}})\rangle$, we get that as $\varepsilon \to 0$,

$$\phi((1+\varepsilon)\hat{\mathbf{x}}) - \phi(\hat{\mathbf{x}}) = \varepsilon\hat{\mathbf{x}}^T\nabla\phi(\hat{\mathbf{x}}) + o(\varepsilon).$$

So the left side of the last expression will be positive, contradicting the assumption that $\hat{\mathbf{x}}$ maximizes ϕ over $\mathcal{K}$.

Proof of Theorem 3.1

To prove the uniqueness of the NPMLE, we calculate the second derivative matrix of ϕ. The diagonal elements are given by (recall that $\Delta_{(m+1)} = 0$)

$$\nabla^2\phi_{ii}(x) = -\left[\frac{1_{[\Delta_{(i)}=1]}}{(x_i - x_{i-1})^2} + \frac{1_{[\Delta_{(i+1)}=1]}}{(x_{i+1} - x_i)^2} + \frac{\delta_{(i)}}{(1-x_i)^2} + \frac{1_{[\Delta_{(i)}=3]}}{x_i^2}\right],$$

for $i = 1, 2, \ldots, s$. The first off-diagonal elements are given by

$$\nabla^2\phi_{i,i-1}(x) = \frac{1_{[\Delta_{(i)}=1]}}{(x_i - x_{i-1})^2}, \quad \text{for} \quad i = 1, 2, \ldots, s,$$

and all the other elements of $\nabla^2\phi(y)$ are 0s.

At the maximizing point y, the second derivative matrix $\nabla^2\phi(y)$ is well defined. Let

$$a_i = \frac{1_{[\Delta_{(i)}=1]}}{(y_i - y_{i-1})^2},$$

$$b_i = \frac{\delta_{(i)}}{(1-y_i)^2},$$

$$c_i = \frac{1_{[\Delta_{(i)}=3]}}{y_i^2},$$

$$d_i = a_{i+1} \ (a_{s+1} = 0),$$

and

$$e_i = a_i + b_i + c_i + d_i$$

for $i = 1, 2, \ldots, s$. Then a_i, b_i, c_i, d_i, and e_i are all nonnegative at y, and the second derivative matrix $-\nabla^2\phi(y)$ is given by

$$-\nabla^2\phi(y) = \begin{bmatrix} e_1 & -d_1 & 0 & & \cdots & 0 \\ -d_1 & e_2 & -d_2 & & \cdots & 0 \\ 0 & -d_2 & e_3 & -d_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ \vdots & \vdots & \vdots & \ddots & -d_{s-1} & \\ 0 & \cdots & -d_{s-1} & & e_s & \end{bmatrix}.$$

We prove that this is a positive definite matrix. Without loss of generality, we may assume in this proof that all the d_i's for $i = 1, 2, \ldots, s-1$ are positive. This is because the matrix would become a block diagonal matrix consisting of two tridiagonal matrices of smaller size if, say, $d_k = 0$ for a k such that $1 < k < s-1$ and no other d_i's were 0. Then, because a block diagonal matrix is positive definite iff all of its submatrices are positive definite, the proof can be applied to each of the component submatrices.

Under the condition that d_i's are all positive, we show that all the determinants J_i of the leading principal minors $M_i = [-\nabla^2\phi_{k,l}(y)]_{k,l=1}^{i}$ of $-\nabla^2\phi(y)$ are positive.

In fact, $J_1 = e_i = a_1 + b_1 + c_1 + d_1 > 0$ and, using $a_2 = d_1$,

$$\begin{aligned} J_2 &= e_2 J_1 - d_1^2 \\ &= (a_1 + b_1 + c_1 + d_1)(a_2 + b_2 + c_2 + d_2) - d_1^2 \\ &= (a_1 + b_1 + c_1 + d_1)(d_1 + b_2 + c_2 + d_2) - d_1^2 \\ &\geq (a_1 + b_1 + c_1)(d_1 + b_2 + c_2 + d_2) > 0. \end{aligned}$$

Now suppose that $J_1, J_2, \ldots, J_{i-1} > 0$ have been proved. Because of the tridiagonal structure of the matrix $-\nabla^2\phi(y)$, we have for $i \geq 3$,

$$J_i = e_i J_{i-1} + (-1)^{2i-1} d_{i-1}^2 J_{i-2} \tag{A.1}$$

$$= (a_i + b_i + c_i + d_i) J_{i-1} - d_{i-1}^2 J_{i-2}. \tag{A.2}$$

Rearranging the last equation by moving the term $d_i J_{i-1}$ to the left side and noting that $a_i = d_{i-1}$, we obtain

$$J_i - d_i J_{i-1} = (a_i + b_i + c_i) J_{i-1} - d_{i-1}^2 J_{i-2} \tag{A.3}$$

$$= d_{i-1}(J_{i-1} - d_{i-1} J_{i-2}) + (b_i + c_i) J_{i-1} \tag{A.4}$$

$$\geq d_{i-1}(J_{i-1} - d_{i-1} J_{i-2}). \tag{A.5}$$

Thus, using (A.2) again in this step,

$$\begin{aligned} J_i &\geq d_{i-1} J_{i-1} - d_{i-1}^2 J_{i-2} \\ &= d_{i-1}[(a_{i-1} + b_{i-1} + c_{i-1} + d_{i-1}) J_{i-2} - d_{i-2}^2 J_{i-3}] \\ &\quad - d_{i-1}^2 J_{i-2} \\ &= d_{i-1}[(a_{i-1} + b_{i-1} + c_{i-1}) J_{i-2} - d_{i-2}^2 J_{i-3}] \\ &\geq d_{i-1} d_{i-2}[J_{i-2} - d_{i-2} J_{i-3}]. \end{aligned}$$

Iterating with (A.5), we obtain

$$\begin{aligned} J_i &\geq d_{i-1} d_{i-2}[J_{i-2} - d_{i-2} J_{i-3}] \\ &\geq \cdots \\ &\geq d_{i-1} d_{i-2} \ldots d_2 [J_2 - d_2 J_1]. \end{aligned}$$

Note that $a_1 + b_1 + c_1$ and $a_2 + b_2 + c_2 > 0$, because $1_{[\Delta_{(i)}=1]} + 1_{[\Delta_{(i)}=2]} + 1_{[\Delta_{(i)}=3]} = 1$ for any $i > 1$. Thus

$$\begin{aligned} J_2 - d_2 J_1 &= (a_1 + b_1 + c_1 + d_1)(a_2 + b_2 + c_2 + d_2) \\ &\quad - d_1^2 - d_2(a_1 + b_1 + c_1 + d_1) \\ &= (a_1 + b_1 + c_1 + d_1)(a_2 + b_2 + c_2) - d_1^2 \\ &= (a_1 + b_1 + c_1)(a_2 + b_2 + c_2) + d_1(b_2 + c_2) > 0. \end{aligned}$$

Hence $J_i > 0$. By induction, all of the leading principal minors are positive. Thus we conclude that $-\nabla^2\phi(y)$ is a positive definite matrix, and $\nabla^2\phi(y)$ is a negative definite matrix, and hence the maximizing point y is unique.

Proof of Theorem 3.2

Suppose $\mathbf{x}^* = \hat{\mathbf{x}} \equiv \arg\max_{x \in \mathcal{K}} \phi(\mathbf{x})$. Note that the derivative vector of $J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}})$ with respect to $\mathbf{x}$ is given by $-\boldsymbol{\Sigma}[\mathbf{x} - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}\nabla\phi(\hat{\mathbf{x}})]$, and we have

$$\begin{cases} -\mathbf{x}^{*T}\boldsymbol{\Sigma}[\mathbf{x}^* - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}\nabla\phi(\hat{\mathbf{x}})] = \langle \hat{\mathbf{x}}, \nabla\phi(\hat{\mathbf{x}})\rangle = 0, \\ -\mathbf{x}^{T}\boldsymbol{\Sigma}[\mathbf{x}^* - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}\nabla\phi(\hat{\mathbf{x}})] = \langle \mathbf{x}, \nabla\phi(\hat{\mathbf{x}})\rangle \leq 0 \end{cases}$$

for any $\mathbf{x} \in \mathcal{K}$. Note that $J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}})$ is a concave function continuously differentiable defined on $\mathcal{K}$. By Lemma 3.1, $\mathbf{x}^*$ maximizes $J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}})$ over $\mathcal{K}$.

On the other hand, suppose that $\mathbf{x}^*$ maximizes $J_{\boldsymbol{\Sigma}}(\mathbf{x}, \hat{\mathbf{x}})$ over $\mathcal{K}$; then we have

$$\begin{cases} -\mathbf{x}^{*T}\boldsymbol{\Sigma}[\mathbf{x}^* - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}\nabla\phi(\hat{\mathbf{x}})] = 0, \\ -\mathbf{x}^{T}\boldsymbol{\Sigma}[\mathbf{x}^* - \hat{\mathbf{x}} - \boldsymbol{\Sigma}^{-1}\nabla\phi(\hat{\mathbf{x}})] \leq 0 \end{cases} \tag{A.6}$$

for any $\mathbf{x} \in \mathcal{K}$ by Lemma 3.1.

From the equality in (A.6), we have

$$\mathbf{x}^{*T}\boldsymbol{\Sigma}[\mathbf{x}^* - \hat{\mathbf{x}}] = \mathbf{x}^{*T}\nabla\phi(\hat{\mathbf{x}}). \tag{A.7}$$

From the inequality in (A.6), we have

$$-\mathbf{x}^{T}\boldsymbol{\Sigma}(\mathbf{x}^* - \hat{\mathbf{x}}) \leq -\mathbf{x}^{T}\nabla\phi(\hat{\mathbf{x}}). \tag{A.8}$$

Adding (A.7) and (A.8) gives

$$(\mathbf{x}^* - \mathbf{x})^T \boldsymbol{\Sigma}(\mathbf{x}^* - \hat{\mathbf{x}}) \leq (\mathbf{x}^* - \mathbf{x})^T \nabla\phi(\hat{\mathbf{x}}).$$

Setting $\mathbf{x} = \hat{\mathbf{x}}$ in the foregoing inequality leads to

$$(\mathbf{x}^* - \hat{\mathbf{x}})^T \boldsymbol{\Sigma}(\mathbf{x}^* - \hat{\mathbf{x}}) \leq \mathbf{x}^* \nabla\phi(\hat{\mathbf{x}}) \leq 0.$$

The last inequality is obtained by Lemma 3.1, because $\mathbf{x}^*$ is a point in $\mathcal{K}$ and $\hat{\mathbf{x}}$ maximizes ϕ over $\mathcal{K}$. But this inequality together with the positive definiteness of $\boldsymbol{\Sigma}$ implies that $\mathbf{x}^* = \hat{\mathbf{x}}$.

Proof of Lemma 3.2 (Jongbloed 1995a, p. 16)

Let $\varphi(\lambda) = \phi(\mathbf{x} + \lambda(A(\mathbf{x}) - \mathbf{x}))$ for a fixed $\mathbf{x} \in \mathcal{K} - \{\hat{\mathbf{x}}\}$. It suffices to show that the right derivative of φ at 0,

$$\varphi'(0) = (A(\mathbf{x}) - \mathbf{x})^T \nabla\phi(\mathbf{x})$$

is strictly positive.

From the optimality conditions in (11) and (12) applied to $J_{\boldsymbol{\Sigma}(\mathbf{x})}(\mathbf{z}, \mathbf{x})$, we have

$$\langle A(\mathbf{x}), -\boldsymbol{\Sigma}(\mathbf{x})(A(\mathbf{x}) - \mathbf{x}) + \nabla\phi(\mathbf{x})\rangle = 0 \tag{A.9}$$

and

$$\langle \mathbf{x}, -\boldsymbol{\Sigma}(\mathbf{x})(A(\mathbf{x}) - \mathbf{x}) + \nabla\phi(\mathbf{x})\rangle \leq 0. \tag{A.10}$$

Subtracting (A.10) from (A.9), we have

$$\begin{aligned}\langle A(\mathbf{x}) - \mathbf{x}, -\boldsymbol{\Sigma}(\mathbf{x})(A(\mathbf{x}) - \mathbf{x}) + \nabla\phi(\mathbf{x})\rangle \\ = -\langle A(\mathbf{x}) - \mathbf{x}, \boldsymbol{\Sigma}(\mathbf{x})(A(\mathbf{x}) - \mathbf{x})\rangle + \varphi'(0) \geq 0.\end{aligned}$$

Because $\boldsymbol{\Sigma}(\mathbf{x})$ is positive definite and $\mathbf{x} \neq \hat{\mathbf{x}}$, it follows that the first term on the left side of the last expression is strictly positive, and hence $\varphi'(0) > 0$.

Proof of Lemma 4.1

From Lemma 3.2 and the definition of N in (16), it follows that the mapping N is well defined and the function ϕ is an ascent function: For all $\mathbf{x} \neq \hat{\mathbf{x}}$ and for all $\mathbf{z} \in N(\mathbf{x})$, $\phi(\mathbf{z}) > \phi(\mathbf{x})$. From this, it follows that

$$\{\mathbf{x}^{(k)} : k \geq 0\} \subset K$$

where K is as defined in (20). From the continuity of the function ϕ on the set $\{\mathbf{x} : \phi(\mathbf{x}) > -\infty\}$ and the assumption that $\phi(\mathbf{x}^{(0)}) > -\infty$, it follows that the set K is compact.

We now show that N is closed at each $\mathbf{x} \in K - \{\hat{\mathbf{x}}\}$. Fix $\mathbf{x} \in K - \{\hat{\mathbf{x}}\}$ and let a sequence $\{\mathbf{x}_k\} \in K$ be such that $\mathbf{x}_k \to \mathbf{x}$. Let $\mathbf{z}_k \in N(\mathbf{x}_k)$ be such that $\mathbf{z}_k \to \mathbf{z}$ for some $\mathbf{z} \in K$. We need to prove that $\mathbf{z} \in N(\mathbf{x})$.

Note that the continuity of the derivative vector $\nabla\phi(\mathbf{x})$ for $\mathbf{x} \in K$ and the continuity of the mapping $\mathbf{x} \mapsto \boldsymbol{\Sigma}(\mathbf{x})$ on K imply that the function $J_{\boldsymbol{\Sigma}(\mathbf{x}_k)}(\mathbf{w}, \mathbf{x}_k) \to J_{\boldsymbol{\Sigma}(\mathbf{x})}(\mathbf{w}, \mathbf{x})$ locally uniformly in $\mathbf{w}$ near $\mathbf{x}$ as $k \to \infty$. Hence the argmax of $J_{\boldsymbol{\Sigma}(\mathbf{x}_k)}(\mathbf{w}, \mathbf{x}_k)$ approaches the argmax of $J_{\boldsymbol{\Sigma}(\mathbf{x})}(\mathbf{w}, \mathbf{x})$ as $k \to \infty$. This implies that

$$A(\mathbf{x}_k) \to A(\mathbf{x}) \tag{A.11}$$

as $k \to \infty$. But $\mathbf{z}_k$ is within the segment between $\mathbf{x}_k$ and $A(\mathbf{x}_k)$ by the definition of $\mathbf{z}_k$: $\mathbf{z}_k \in \text{seg}(\mathbf{x}_k, A(\mathbf{x}_k))$. Hence $\mathbf{z}_k \to \mathbf{z}$ implies the point $\mathbf{z} \in \text{seg}(\mathbf{x}, A(\mathbf{x}))$ necessarily. Now consider the two different situations that can occur.

The first situation is that

$$\phi(A(\mathbf{x}_k)) \geq \phi(\mathbf{x}_k) + (1 - \varepsilon)\nabla\phi(\mathbf{x}_k)^T (A(\mathbf{x}_k) - \mathbf{x}_k)$$

for infinitely many values of k. Let k go to infinity along a subsequence k_j where this inequality holds; we get from (A.11) that

$$\phi(A(\mathbf{x})) \geq \phi(\mathbf{x}) + (1 - \varepsilon)\nabla\phi(\mathbf{x})^T (A(\mathbf{x}) - \mathbf{x}).$$

It is thus trivially true $N(\mathbf{x}) = A(\mathbf{x})$ when the above inequality holds strictly. When the foregoing inequality is an equality, we still have $N(\mathbf{x}) = A(\mathbf{x})$, because $(1 - \varepsilon) > \varepsilon$, $\nabla\phi(\mathbf{x})^T (A(\mathbf{x}) - \mathbf{x}) > 0$, and

$$\begin{aligned}\phi(A(\mathbf{x})) &= \phi(\mathbf{x}) + (1 - \varepsilon)\nabla\phi(\mathbf{x})^T (A(\mathbf{x}) - \mathbf{x}) \\ &> \phi(\mathbf{x}) + \varepsilon\nabla\phi(\mathbf{x})^T (A(\mathbf{x}) - \mathbf{x}).\end{aligned}$$

Moreover, along the same subsequence, it follows from the definition of N that $\mathbf{z}_{k_j} \in A(\mathbf{x}_{k_j})$. Therefore, $\mathbf{z}_{k_j} \to A(\mathbf{x})$ along the same subsequence by the continuity of A. This shows that $\mathbf{z} \in A(\mathbf{x}) = N(\mathbf{x})$, as was to be proved.

The other possibility is that for all k sufficiently large,

$$\phi(A(\mathbf{x}_k)) < \phi(\mathbf{x}_k) + (1 - \varepsilon)\nabla\phi(\mathbf{x}_k)^T (A(\mathbf{x}_k) - \mathbf{x}_k).$$

Let $k \to \infty$ and use (A.11); it follows that

$$\phi(A(\mathbf{x})) \leq \phi(\mathbf{x}) + (1 - \varepsilon)\nabla\phi(\mathbf{x})^T (A(\mathbf{x}) - \mathbf{x}).$$

Therefore, by the definition of N and the fact that $\mathbf{z} \in \text{seg}(\mathbf{x}, A(\mathbf{x}))$, we have $\mathbf{z} \in N(\mathbf{x})$ if we have

$$\begin{aligned}\phi(\mathbf{x}) + \varepsilon\nabla\phi(\mathbf{x})^T (\mathbf{z} - \mathbf{x}) \leq \phi(\mathbf{z}) \\ \leq \phi(\mathbf{x}) + (1 - \varepsilon)\nabla\phi(\mathbf{x})^T (\mathbf{z} - \mathbf{x}).\end{aligned}$$

But this follows from the fact that $\mathbf{z}_k \in A(\mathbf{x}_k)$ for all k sufficiently large and

$$\begin{aligned}\phi(\mathbf{x}_k) + \varepsilon\nabla\phi(\mathbf{x}_k)^T (\mathbf{z}_k - \mathbf{x}_k) \leq \phi(\mathbf{z}_k) \\ \leq \phi(\mathbf{x}_k) + (1 - \varepsilon)\nabla\phi(\mathbf{x}_k)^T (\mathbf{z}_k - \mathbf{x}_k).\end{aligned}$$

Proof of Theorem 4.2

Take $\mathbf{X}$ to be the feasible set $\mathbf{C}_\mathbf{x}$ as given in (2). It is a nonempty closed set in $\mathbb{R}^n$. The solution set is taken as $\Omega = \{\hat{\mathbf{x}}\} \subset \mathbf{C}_\mathbf{x}$. Take $\alpha = -\phi$.

Take $\mathbf{B}(\cdot)$ and $\mathbf{C}(\cdot)$ in Theorem 4.1 to be the algorithmic mappings $N(\cdot)$ and $M(\cdot)$ of the modified ICM and the EM algorithm. Then by inequality (19), we have $\alpha(\mathbf{z}) = -\phi(\mathbf{z}) \leq -\phi(\mathbf{x}) = \alpha(\mathbf{x})$ for any $\mathbf{z} \in M(\mathbf{x})$ and $\mathbf{x} \in \mathbf{C}_\mathbf{x}$. By Lemma 4.1, $\mathbf{B} = N$ is closed on the complement of Ω, and we have $\alpha(\mathbf{z}) = -\phi(\mathbf{z}) < -\phi(\mathbf{x}) = \alpha(\mathbf{x})$ for each $\mathbf{z} \in N(\mathbf{x}) = \mathbf{B}(\mathbf{x})$ and $\mathbf{x} \notin \Omega$. Finally, because ϕ is continuous and $\mathbf{x}^{(0)}$ is such that $\phi(\mathbf{x}^{(0)}) > -\infty$, the set $\{\mathbf{x} : \alpha(\mathbf{x}) \leq \alpha(\mathbf{x}^{(0)})\}$ is compact. Hence by Theorem 4.1, either the hybrid algorithm stops in a finite number of steps with an iterate $\mathbf{x}^{(k)} = \hat{\mathbf{x}}$ or it generates an infinite sequence $\{\mathbf{x}^{(k)}\}$ such that $\hat{\mathbf{x}}$ is the only accumulation point of $\{\mathbf{x}^{(k)}\}$. Hence $\mathbf{x}^{(k)}$ converges to $\hat{\mathbf{x}}$.

Proof of Proposition 4.1

Suppose that $M(\hat{\mathbf{x}}) \neq \hat{\mathbf{x}}$. Then the uniqueness of $\hat{\mathbf{x}}$ implies that $\phi(\hat{\mathbf{x}}) > \phi(M(\hat{\mathbf{x}}))$, which contradicts the definition of the EM algorithmic mapping.

Similarly, if $MN(\hat{\mathbf{x}}) \neq \hat{\mathbf{x}}$ (or $N(\hat{\mathbf{x}}) \neq \hat{\mathbf{x}}$), then by the definition of the mapping M and the definition of N, we have

$$\phi(M(N(\hat{\mathbf{x}}))) \geq \phi(N(\hat{\mathbf{x}})) > \phi(\hat{\mathbf{x}}),$$

which contradicts $\phi(\hat{\mathbf{x}}) > \phi(M(N(\hat{\mathbf{x}})))$ by the uniqueness of $\hat{\mathbf{x}}$ (or $\phi(\hat{\mathbf{x}}) > \phi(N(\hat{\mathbf{x}}))$).

Proof of Corollary 4.1

Because $\hat{\mathbf{x}}$ is a fixed point of the EM algorithmic mapping $M(\cdot)$, it is a limit point of the EM iterates. Hence $\hat{\mathbf{x}}$ satisfies the self-consistency equations (5).

[Received September 1995. Revised December 1996.]

REFERENCES

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955), "An Empirical Distribution Function for Sampling With Incomplete Information," *Annals of Mathematical Statistics*, 26, 641–647.

Barlow, R. E., Bartholomew, R. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference Under Order Restrictions*. New York: Wiley.

Bazaraa, M. S., Sherali, H. D., and Shetti, C. M. (1993), *Nonlinear Programming, Theory and Algorithms*. New York: Wiley.

Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Chang, M. N. (1990), "Weak Convergence of a Self-Consistent Estimator of the Survival Function With Doubly Censored Data," *The Annals of Statistics*, 18, 391–404.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Efron, B. (1967), "The Two-Sample Problem With Censored Data," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics*, 4, 831–853.

Gill, R. (1989), Non- and Semi-parametric Maximum Likelihood Estimators and the von Mises Method (Part 1)," *Scandinavian Journal of Statistics*, 16, 97–128.

Groeneboom, P. (1996), "Lectures on Inverse Problems (École d'Été de Probabilités de Saint-Flour XXIV-1994)," in *Lecture Notes in Mathematics*, ed. P. Bernard, New York: Springer-Verlag, pp. 67–164.

Groeneboom, P., and Wellner, J. A. (1992), "Information Bounds and Nonparametric Maximum Likelihood Estimation, in *DMV Seminar Band 19*, Basel: Birkhäuser Verlag.

Gu, M. G., and Zhang, C.-H. (1993), "Asymptotic Properties of Self-Consistent Estimators Based on Doubly Censored Data," *The Annals of Statistics*, 21, 611–624.

Jamshidian, M. J., and Jennrich, R. I. (1993), "Conjugate Gradient Acceleration of the EM Algorithm," *Journal of the American Statistical Association*, 88, 221–228.

Jongbloed, G. (1995a), "Three Statistical Inverse Problems," unpublished Ph.D. dissertation, Delft University.

——— (1995b), "The Iterative Convex Minorant Algorithm for Nonparametric Estimation," submitted to *Journal of Computational and Graphical Statistics*.

Laird, N. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–815.

Meilijson, I. (1989), "A Fast Improvement to the EM Algorithm on Its Own Terms," *Journal of the Royal Statistical Society*, Ser. B, 51, 127–138.

Mykland, P. A., and Ren, J.-J. (1996), Algorithm for Computing Self-Consistent and Maximum Likelihood Estimators With Doubly Censored Data," *The Annals of Statistics*, 24, 1740–1764.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order Restricted Statistical Inference*, New York: Wiley.

Tsai, W. Y., and Crowley, J. (1985), "A Large Sample Study of Generalized Maximum Likelihood Estimators From Incomplete Data via Self-Consistency," *The Annals of Statistics*, 13, 1317–1334. Corr. 18, 470.

Turnbull, B. W. (1974), "Nonparametric Estimation of a Survival Ship Function With Doubly Censored Data," *Journal of the American Statistical Association*, 69, 169–173.

——— (1976), "The Empirical Distribution Function With Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society*, Ser. B, 38, 290–295.

Turnbull, B. W., and Weiss, L. (1978), "A Likelihood Ratio Statistic for Testing Goodness of Fit With Randomly Censored Data," *Biometrics*, 34, 367–375.

Van der Vaart, A. W. (1995), "Efficiency of Infinite Dimensional M-Estimators," *Statistica Neerlandica*, 49, 9–30.

Van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Van Eeden, C. (1956), "Maximum Likelihood Estimation of Ordered Probabilities," *Indagationes Mathematicae*, 18, 444–455.

Vardi, Y. (1982), "Nonparametric Estimation in Renewal Processes," *The Annals of Statistics*, 10, 772–785.

——— (1989), "Multiplicative Censoring, Renewal Processes, Deconvolution and Decreasing Density: Nonparametric Estimation," *Biometrika*, 76, 751–761.

Vardi, Y., and Zhang, C.-H. (1992), "Large-Sample Study of Empirical Distributions in a Random-Multiplicative Censoring Model," *The Annals of Statistics*, 20, 1022–1039.

Wellner, J. A., and Zhan, Y. (1996), "Bootstrapping Z-Estimators," technical report, University of Washington, Dept. of Statistics.

Wu, J. F. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.

Zhan, Y. (1996), "Bootstrapping Functional M-Estimators," unpublished Ph.D. dissertation, University of Washington, Dept. of Statistics.

Zhan, Y., and Wellner, J. A. (1995), "Double censoring: Characterization and Computation of the Nonparametric Maximum Likelihood Estimator," Technical Report 292, University of Washington, Dept. of Statistics.

Zhou, M. (1993), "Bootstrapping the Survival Curve Estimator When Data are Doubly Censored," Technical Report 335, University of Kentucky, Dept. of Statistics.