UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS

Technical Report #115

January 1997

Computation of the NPMLE of distribution functions for interval censored and truncated data with applications to the Cox model

Wei Pan
Richard Chappell, Ph.D.

Department of Biostatistics

MADISON, WISCONSIN

Computation of the NPMLE of distribution functions for interval censored and truncated data with applications to the Cox model

Wei Pan, Rick Chappell

Departments of Statistics and Biostatistics, University of Wisconsin-Madison, USA

Abstract

The iterative convex minorant (ICM) algorithm (Groeneboom and Wellner, 1992) is widely believed to be much faster than the EM algorithm in computing nonparametric MLEs of distribution functions for interval censored data. We first extend the ICM to left truncated and interval censored data. Our formulation of the ICM helps to explore its connection with the gradient projection method that is commonly used in the constrained optimization area. Some modifications to the ICM naturally suggest new algorithms. In particular a damped ICM (DICM), a gradient projection based (GP) method, a hybrid of the damped ICM and EM, and a hybrid of the GP and EM are proposed. Simulations were conducted to compare their convergence speeds and accuracies, in which the GP seems very promising. An important application of the GP method is to the Cox proportional hazards model with left truncated and interval censored data. The methodology is illustrated by using the Massachusetts Health Care Panel Study dataset.

Key Words: EM algorithm; Gradient projection; Greatest convex minorant; ICM algorithm; Isotonic regression; Proportional hazards model.

1. Introduction

Interval censored data occur commonly, such as in studies of AIDS and other chronic diseases, in which the the events of interest are never observed exactly but only known to occur within some time intervals. Turnbull (1976) presents an EM algorithm (Dempster, Laird and Rubin, 1977) to compute the nonparametric maximum likelihood estimator (NPMLE) of distribution functions for arbitrarily censored and truncated data. Groeneboom and Wellner (1992, shortened as G&W in the sequel) present an iterative convex minorant (ICM) algorithm to compute the NPMLE of distribution functions for interval censored data. G&W suggest that empirical studies show the ICM is much faster than the EM. There has been a large amount of work attempting acceleration of the EM (Meilijson, 1989; Jamshidian and Jennrich, 1993; also see later references). Suggestions involve

using second-order information, such as in the Newton-Raphson or quasi-Newton type algorithms (Atkinson, 1992; Lange, 1995b). This approach tries to improve the locally linear convergence rate of the EM to the quadratic or super-linear convergence rate. Others (Meng and Rubin, 1992; Rai and Matthews, 1993; Liu and Rubin, 1994; Lange, 1995a) propose to speed up the M step when implementing it is not straightforward. In estimating distribution functions for interval censored data, neither approach may fit. For the first one, since the number of (unknown and to be estimated) parameters is in the order of the number of observations (which is frequently more than several hundred), in practice often we just simply cannot afford the space required for the Hessian matrix or its approximation and the inversion operation incurred on today's ordinary workstations. For the second, we have a simple explicit solution for the M step in the EM algorithm, hence no need to accelerate the M step. However, since the ICM seems faster than the EM, some further analysis of the ICM may help to find some new faster algorithms. Moreover, Zhan and Wellner (1995) show that a hybrid algorithm alternating the ICM step and the EM step is much faster than the ICM or the EM alone in computing the NPMLE of the distribution functions for doubly censored data. There may be still some room for improving the speed along this direction.

In addition to censoring, truncation may also arise, such as in the Massachusetts Health Care Panel Study (Branch and Ku, 1989; Chappell, 1991). In these studies a baseline examination was conducted, then after some fixed time the first followup took place, then second and third, *etc.* followups (if any) after intervals of several years each. Left truncation is introduced because we know for sure the event of interest (such as death) can only happen after the entries of the subjects (who may have different ages). If the event happens, we can only know it lies between two examination times. Otherwise, it will happen after the end of the last followup. This introduces interval and right censoring. Since there is still no other well-known method available except the EM in computing the NPMLE of distribution functions for left truncated and interval censored data, it is interesting to discuss the computational methods in this more general setting.

The computation of NPMLEs of distribution functions is essentially a constrained maximization problem: the log-likelihood is maximized with the constraint that the maximizing point is a proper distribution (or subdistribution). Hence it is natural to apply the existing constrained optimization techniques, among which three are commonly used (Polak, 1971). The first is the Iterative Quadratic Programming, which iteratively approximates the object function by a quadratic form

and applies a quadratic programming algorithm to maximize the quadratic form. It is very like the Newton-Raphson algorithm in unconstrained optimization. The drawback is the Hessian matrix of the object function is needed, which may be impractical in our case. The second is the class of the Penalty Methods, which incoporates the constraint into the objective function via a penalty term. The problem is that the coefficient of penalty term must approach some limit, such as 0 or ∞, hence it may depend on user's interaction to stop the algorithm. The last is the Gradient Projection Method, which projects the new point along the gradient of the objective function into the constraint region. In many cases, it is hard to compute a projection of a point to another (convex) region. But in our case, the constrained region is simple enough for the projection to be efficiently accomplished by some well-known isotonic regression algorithms.

In this paper, we first derive an ICM algorithm for left truncated and interval censored data, in which we show its connection with the gradient projection (GP) method. Then we propose a GP algorithm and a hybrid algorithm of the GP and EM and a hybrid of the ICM and EM. Simulations were conducted to show their performances in terms of convergence speed and accuracy. The GP method is also further applied to the Cox proportional hazards model with left truncated and interval censored data with some simulations. The MHCPS is taken as an example to illustrate this methodology.

2. Extension of ICM to truncated data and its modifications

Let (X_1, T_1, U_1, V_1), ...,(X_n, T_n, U_n, V_n) be a sample of random variables in R_+^4, where X_i is a non-negative random variable whose distribution function F_0 is what we are interested in estimating, and where T_i, U_i and V_i are truncation and censoring random variables respectively. The only observations available are $(T_i, U_i, V_i, \delta_{1,i} = I(T_i \le X_i \le U_i), \delta_{2,i} = I(U_i < X_i \le V_i), \delta_{3,i} = I(X_i > V_i))$, $i = 1, 2, ..., n$. Notice that it is impossible to observe that $X_i < T_i$, i.e. $\delta_{1,i} + \delta_{2,i} + \delta_{3,i} = 1$ for $i = 1, 2, ..., n$.

The log-likelihood divided by n can be written as:

$$\psi(F) := \int_{R_+^4} \phi_F(x, t, u, v) dP_n(x, t, u, v),$$

where P_n is the empirical probability measure of the (X_i, T_i, U_i, V_i), $1 \le i \le n$, and

$$\phi_F(x, t, u, v) := I(t \le x \le u) \log\{F(u) - F(t)\}$$

$$+ \quad I(u < x \le v)\log\{F(v) - F(u)\} + I(x > v)\log\{1 - F(v)\} - \log\{1 - F(t)\}.$$

The *nonparametric maximum likelihood estimator* (NPMLE) F_n of F_0 is a right-continuous function F maximizing $\psi(F)$. Turnbull (1976) shows that F_n can only have jumps between the order statistics τ_i, $i = 1, ..., k$, of T_i, U_i and V_i, $1 \le i \le n$. However, the probability distribution of F_n within some of the intervals is nonidentifiable. This nonidentifiability will not influence the computation of F_n by the EM algorithm. To facilitate the computation of F_n by the ICM, G&W define the NPMLE F_n as piecewise constant except at the order statistics $\{\tau_i\}$, and moreover F_n may be less than 1 at the largest order statistic τ_k. These two different definitions seem to be non-critical. In the sequel, we adopt the second one except when we are dealing with the EM.

Turnbull (1976) also suggests how to reduce the number of the order statistics $\{\tau_i\}$ on which F_n may have jumps. Notice that Frydman (1994) corrects a problem in this reduction when applied to truncated data. In the sequel, we suppose this correction is in use.

Now approaching as in G&W, we define the processes: for $t \ge 0$,

$$\begin{aligned}
W_F(t) = & \int_{y\in[0,t],t'\le x\le y} \{F(y) - F(t')\}^{-1} dP_n(x, t', y, v) \\
& - \int_{y\in[0,t],y\le x\le u} \{F(u) - F(y)\}^{-1} dP_n(x, y, u, v) + \int_{y\in[0,t],u<x\le y} \{F(y) - F(u)\}^{-1} dP_n(x, t', u, y) \\
& - \int_{y\in[0,t],y<x\le v} \{F(v) - F(y)\}^{-1} dP_n(x, t', y, v) - \int_{y\in[0,t],x>y} \{1 - F(y)\}^{-1} dP_n(x, t', u, y) \\
& + \int_{y\in[0,t]} \{1 - F(y)\}^{-1} dP_n(x, y, u, v),
\end{aligned}$$

$$\begin{aligned}
G_F(t) = & \int_{y\in[0,t],t'\le x\le y} \{F(y) - F(t')\}^{-2} dP_n(x, t', y, v) \\
& + \int_{y\in[0,t],y\le x\le u} \{F(u) - F(y)\}^{-2} dP_n(x, y, u, v) + \int_{y\in[0,t],u<x\le y} \{F(y) - F(u)\}^{-2} dP_n(x, t', u, y) \\
& + \int_{y\in[0,t],y<x\le v} \{F(v) - F(y)\}^{-2} dP_n(x, t', y, v) + \int_{y\in[0,t],x>y} \{1 - F(y)\}^{-2} dP_n(x, t', u, y) \\
& - \int_{y\in[0,t]} \{1 - F(y)\}^{-2} dP_n(x, y, u, v),
\end{aligned}$$

and

$$Q_F(t) = W_F(t) + \int_{y\in[0,t]} F(y) dG_F(y).$$

Because of truncation, we have a trouble here: $G_F(t)$ may be no longer increasing. If it is increasing as in the interval-censoring (without truncation) case, we may as in G&W characterize

the NPMLE of F_0 as the left derivative of the convex minorant of the "cumulative sum diagram":

$$P_0 = (0,0), \quad P_i = \left(G_{F_n}(\tau_{(i)}), Q_{F_n}(\tau_{(i)})\right),$$

where $i = 1, 2, ..., k$. By definition, the left derivative h_n of the convex minorant of the cumulative sum diagram is

$$h_n(\tau_i) = \frac{Q_{F_n}(\tau_i) - Q_{F_n}(\tau_{i-1})}{G_{F_n}(\tau_i) - G_{F_n}(\tau_{i-1})} = F_n(\tau_i) + \frac{\Delta W_{F_n}(\tau_i)}{\Delta G_{F_n}(\tau_i)},$$

where $\Delta W_{F_n}(\tau_i) = W_{F_n}(\tau_i) - W_{F_n}(\tau_i^-)$ and $\Delta G_{F_n}(\tau_i) = G_{F_n}(\tau_i) - G_{F_n}(\tau_i^-)$, $i = 1, ..., k$. This naturally suggests an iterative algorithm to compute the NPMLE F_n. Now we derive it formally.

Since $\Delta G_{F_n}(\tau_i)$ may not be always positive with truncated data, we define:

$$\Delta \tilde{G}_F(\tau_i) = \begin{cases} \Delta G_F(\tau_i), & \text{if } \Delta G_F(\tau_i) > 0 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

for $i = 1, ..., k$.

For an iterative algorithm, suppose $F^{(m)}$ is the estimate from the mth iteration in computing the NPMLE. As in G&W, consider a *partial* second order approximation to $\psi(F) - \psi(F^{(m)})$:

$$\begin{aligned} &\int \left(F(t) - F^{(m)}(t)\right) dW_{F^{(m)}}(t) - \frac{1}{2}\int \left(F(t) - F^{(m)}(t)\right)^2 d\tilde{G}_{F^{(m)}}(t) \\ &= \sum_{i=1}^{k} \left[\left(F(\tau_i) - F^{(m)}(\tau_i)\right) \frac{\Delta W_{F^{(m)}}(\tau_i)}{\Delta \tilde{G}_{F^{(m)}}(\tau_i)} - \frac{1}{2}\left(F(\tau_i) - F^{(m)}(\tau_i)\right)^2\right] \Delta \tilde{G}_{F^{(m)}}(\tau_i) \\ &= \sum_{i=1}^{k} \left[\frac{1}{2}F(\tau_i)^2 + \left(F^{(m)}(\tau_i) + \frac{\Delta W_{F^{(m)}}(\tau_i)}{\Delta \tilde{G}_{F^{(m)}}(\tau_i)} - F\right) F(\tau_i) + C_i\right] \Delta \tilde{G}_{F^{(m)}}(\tau_i), \end{aligned} \tag{2}$$

where C_i is independent of F. By Theorem 1.5.1 of Robertson, Wright and Dykstra (1988, page 31), if we take $\Phi(x) = x^2/2$ we immediately have that (2) is maximized uniquely by the solution of the isotonic regression with $(\Delta \tilde{G}_{F^{(m)}}(\tau_i), F^{(m)}(\tau_i) + \Delta W_{F^{(m)}}(\tau_i)/\Delta \tilde{G}_{F^{(m)}}(\tau_i))$, $i = 1, 2, ..., k$, which is readily available by some well-known algorithms such as the pooled-adjacent-violators algorithm (PAVA) (Robertson *et al.*, 1988).

Without ambiguity we will generally denote v as a (column) vector with components $v_i, i = 1, ..., k$. All the arithmetic operations for vectors are component-wise unless specified otherwise. It is easy to show that $\Delta W_{F^{(m)}}$ is the first-order derivative of the log-likelihood ψ at the current estimate $F^{(m)}$, and $\Delta \tilde{G}_{F^{(m)}}$ is the vector whose components approximate the diagonal elements of

the Hessian matrix of $-\psi$ at $F^{(m)}$. Our proposed ICM algorithm for left truncated and interval censored data can be formulated as:

$$F^{(m+1)} = \mathbf{Proj}(F^{(m)} + [\Delta\tilde{G}_{F^{(m)}}]^{-1}\Delta W_{F^{(m)}}),$$

and the projection is defined by

$$\mathbf{Proj}(y) := \arg\min_x \{\sum_{i=1}^{k}(y_i - x_i)^2\Delta\tilde{G}_{x,i} : 0 \le x_1 \le x_2 \le ... \le x_k \le 1\},$$

which can be efficiently accomplished by an isotonic regression algorithm such as the PAVA mentioned above. Notice that if there is no truncation, $\Delta\tilde{G}_F = \Delta G_F > 0$ and hence the above algorithm reduces to the usual ICM for interval censored data.

This characterization makes it clear that at each iteration step the new estimate is obtained via modifying the current estimate by its gradient and part of its Hessian, and making sure that the new estimate is a proper distribution function. We follow Aragon and Eberly (1992) to further define a *damped iterative convex minorant algorithm* (DICM) as

$$F^{(m+1)} = \mathbf{Proj}(F^{(m)} + \alpha_j[\Delta\tilde{G}_{F^{(m)}}]^{-1}\Delta W_{F^{(m)}}),$$

where α_k is chosen by

$$\alpha_j = \max\left\{(1/2)^i : i \ge 0, \psi(F^{(m+1)}) > \psi(F^{(m)})\right\},$$

and the projection is defined as in the ICM. The control of the stepsize α_j can be taken more stringently, for instance as the Amiju stepsize.

The appeal of the DICM is its possible global convergence and increase in log-likelihood at each step, though Zhan and Wellner (1995) point out a problem in the proof of the global convergence in Aragon and Eberly (1992). On the other hand, G&W (page 70) suggest using "buffers" to keep the estimate as a proper distribution and to prevent the likelihood from becoming 0. However, they do not say what to do if the new estimate does make the likelihood 0. We feel that the damped version is cleaner from programming point. Moreover, keeping the log-likelihood increasing is not only a way to prevent it from becoming 0, but also a desirable property of its own. (Notice that one attractive property of the EM algorithm is its increasing log-likelihood in each iteration.) Hence, from now on we will only discuss the damped version of the ICM.

If we approximate $\Delta G_{F^{(m)}}$ further by vector $\mathbf{1}$, we have the following gradient projection-like method:

$$F^{(m+1)} = \mathbf{Proj}(F^{(m)} + \alpha_j \Delta W_{F^{(m)}}),$$

$$\alpha_j = \max\left\{(1/2)^i : i \geq 0, \psi(F^{(m+1)}) > \psi(F^{(m)})\right\},$$

and the projection is defined by

$$\mathbf{Proj}(y) := \arg\min_x \{\sum_{i=1}^{k}(y_i - x_i)^2 : 0 \leq x_1 \leq x_2 \leq ... \leq x_k \leq 1\},$$

which can be also efficiently performed by an isotonic regression algorithm such as the PAVA.

In the above algorithm we have to perform a projection operation for each tried stepsize α_k, which is relatively time-consuming. In our gradient projection (GP) method (see also Polak, 1971; Bertsekas, 1976), the projection operation is performed only once at each step:

$$dF^{(m+1)} = \mathbf{Proj}(F^{(m)} + \Delta W_{F^{(m)}}) - F^{(m)}, \quad F^{(m+1)} = F^{(m)} + \alpha_j dF^{(m+1)},$$

$$\alpha_j = \max\left\{(1/2)^i : i \geq 0, \psi(F^{(m+1)}) > \psi(F^{(m)})\right\}.$$

All of the above parametrizations are in terms of the distribution function $F(x)$. We can also use the cumulative hazard function $H(x) = -\log\{1 - F(x)\}$, which is more convenient in the Cox proportional hazards model in particular (Huang 1994; Huang and Wellner, 1995). Our GP method as reported in the later sections (including in Section 4) is actually implemeted this way.

We use the GP method for its simplicity and generality. In our current situation, we can prove that the log-likelihood is concave and bounded above in the constraint region. Hence the GP method should have desirable convergence properties. Moreover, the GP method has already been successfully applied to compute the monotone MLE (the nonparametric MLE under a monotone hazard assumption) of distribution functions for left truncated and interval censored data (Pan and Chappell, 1996).

Other modifications to the ICM are also possible. One heuristic is that the second order information may not be very helpful when it is still far away from the solution. Therefore, we may selectively use the second order approximation of ψ whenever it is likely to be close to the solution point, such as when the log-likelihood increment from the previous step is small. Another interesting approach is to approximate the Hessian matrix more accurately but not use its full form.

On the other hand, the EM is well-known to be very slow. But as noted by Redner and Walker (1984) in studying mixture problems, it can gain 95% *total* log-likelihood increment in as small as first 5 steps. Zhan and Wellner (1995) propose a hybrid of the ICM and the EM which is shown from their simulations to be much faster than either algorithm alone for doubly censored data. This motivated us to explore some hybrid algorithms of the (modified) ICM and EM for interval censored data. Among them, the hybrid algorithm GP-EM which alternates a GP step and an EM step, and the DICM-EM which alternates a DICM step and an EM step will be discussed. In the next section, we show our simulation results.

3. Simulation for estimating distributions

Two sets of simulations were conducted for interval censored data with or without left truncation respectively. All algorithms (including those in the next section) are implemented in the C language and tested on a Sun Sparc10 running SunOS 4.1.3. The accuracy of an algorithm is evaluated by its maximized log-likelihood while its speed by both the CPU time and the iteration number. Notice that optimizing the code may decrease the CPU time but should not change the iteration number. All starting values for distribution functions were taken to be discrete uniform. The convergence criterion was that the log-likelihood increment is less than 0.00001. For simplicity, we suppose there is only one followup after the baseline examination. To mimic the pattern of the MHCPS and other panel surveys the time interval between the two examinations is taken to be constant, len. Denote the baseline time and the event time of interest as T_i and X_i respectively. For interval censored data without truncation, a random sample is generated by repeating the following steps for $i = 1$ to a fixed number of times:

Step 1. X_i and T_i are generated independently by some specified distributions (from Splus).

Step 2. If $X_i < T_i$, we get an left-censored observation in $[0, T_i)$; if $T_i \le X_i \le T_i + len$, we get one lying in $[T_i, T_i + len]$; else it is right-censored at $T_i + len$.

We take $X_i \sim Gamma(2, 1)$, $T_i \sim U(0, 4)$ (i.e. continuous uniform between 0 and 4) and $len = 0.5$. The results are shown in Table 1. It is verified again that the DICM and the DICM-EM are much faster than others for interval censored data, though the DICM-EM is not faster than the DICM. The third fastest is the GP.

We point out here that there may be some problem in the example given by Aragon and Eberly

Table 1: Mean and standard deviation (in parentheses) of convergence speed (iterations and CPU time in seconds needed) and the maximized log-likelihood value without truncation, $n = 100$, and 100 replications.

	EM	DICM	GP	DICM-EM	GP-EM
CPU time	32.3 (6.7)	0.3 (0.1)	5.5 (6.1)	0.7 (0.1)	21.6 (6.4)
Iterations	821 (139)	9 (6)	197 (207)	16 (3)	533 (151)
log-likelihood	-134.103	-134.737	-134.138	134.100	-134.101
	(9.969)	(9.973)	(9.963)	(9.969)	(9.969)

(1992). According to the data given (i.e. the AIDS data from the U.S. Air Force, Table 1 on page 137 in Aragon and Eberly, 1992), it is impossible for the NPMLE of the distribution function to have a jump after 26.43. However, in their Figure 1 it does have a jump near 40. We tried with the EM, the DICM and the GP, and they all yielded similar results (in terms of survival function, which is simply 1 minus the cumulative distribution function) as shown in Figure 1.

For left truncated and interval censored data, we can only observe X_i in either $[T_i, T_i + len]$ or $(T_i + len, \infty)$ if and only if $X_i \geq T_i$. A random sample is generated by repeating the following steps for $i = 1$ to a fixed number of times:

Step 1. X_i and T_i are generated independently by some specified distributions (from Splus).

Step 2. If $X_i < T_i$, this observation is ignored. Otherwise, if $X_i \leq T_i + len$, we get an observation lying in $[T_i, T_i + len]$; else it is right-censored at $T_i + len$.

We take $X_i \sim Gamma(2, 1)$ and $T_i \sim U(0, 4)$. Two sample sizes were used, 100 and 1000 approximately. (Notice that the sample size is random due to truncation.) The results are shown in Tables 2 and 3. The most surprising finding is that the DICM and the DICM-EM may be slower than the GP, and they may also prematurely stop. Either the EM or the GP-EM is also much slower than the GP. Compared with Table 1, the EM also slows down. One important reason is that with truncation the imputation (i.e. the E-step) is more time-consuming in the EM. In contrast, it seems that the GP is faster with truncated data than without truncation. As the sample size increases to 1000, the EM and the GP-EM are so slow that we give up trying 100 replications. Also notice the unstable performance by two hybrid algorithms.

Table 2: Mean and standard deviation (in parentheses) of convergence speed (iterations and CPU time in seconds needed) and the maximized log-likelihood value with truncation, $n \approx 100$, and 100 replications.

	EM	DICM	GP	DICM-EM	GP-EM
CPU time	92.1 (119.1)	4.9 (10.5)	0.9 (0.6)	11.9 (25.8)	20.9 (31.9)
Iterations	4482 (5655)	529 (1140)	114 (67)	599 (1262)	941 (1370)
log-likelihood	-37.0 (5.9)	-36.8 (5.9)	-36.9 (5.9)	-37.1 (6.1)	-39.1 (7.5)

Table 3: Mean and standard deviation (in parentheses) of convergence speed (iterations and CPU time in seconds needed) and the maximized log-likelihood value with truncation, $n \approx 1000$, and 20 replications.

	EM	DICM	GP	DICM-EM	GP-EM
CPU time	14603.7	730.1	171.6	4500.8	7029.3
	(7935.7)	(1857.6)	(140.8)	(7690.8)	(9024.8)
Iterations	7090 (3977)	1136 (2952)	368 (303)	3577 (6089)	5663 (7386)
log-likelihood	-477.8 (17.9)	-486.2 (34.3)	-476.9 (19.2)	-479.1 (22.0)	-476.6 (19.4)

4. Applications to Cox proportional hazards model

The gradient projection method can be widely applied. Here we illustrate its application to the Cox Proportional Hazards Model. With interval censored data, there are three main approaches: 1) (full) likelihood by maximizing the full likelihood jointly with respect to the baseline hazards and the regression coefficients (Finkelstein, 1986; Alioum and Commenges, 1996); 2) profile likelihood by maximizing the likelihood with fixed regression coefficients, which is only feasible for low dimensional covariate vector (Huang and Wellner, 1993, 1995); 3) marginal likelihood (as in right-censored data case where it is very simple), which however has to recourse to some complex Monte Carlo methods (Sinha, Tanner and Hall, 1994, by the Monte Carlo EM; Satten, 1996, by the Markov chain Monte Carlo and stochastic approximation) since there is no explicit formula for the marginal likelihood available with interval censored data.

The gradient projection method can be directly applied to jointly maximize the log-likelihood with respect to the baseline hazards and the regression coefficients. But computationally this is different from Finkelstein (1986) and Alioum and Commenges (1996), where the combination of the Newton-Raphson method and the steepest descent method is adopted after reparametrization. As pointed out earlier, it is more likely too expensive to use the Hessian matrix for its large dimension on ordinary workstations. On the other hand, it seems that the projection step in the GP method speeds up the convergence. We show in this section that the GP method is a competitive alternative that is general and fast enough when applied in this setting.

The Cox proportional hazards model is the most widely used one in survival analysis. It assumes that the hazard function $h(x|Z)$ with covariate Z is proportional to an unknown baseline hazard $h_0(x)$:

$$h(x|Z) = h_0(x)\exp(Z\beta),$$

or formulated in terms of distribution functions:

$$1 - F(x|Z) = [1 - F_0(x)]^{\exp(Z\beta)},$$

where F_0 is the corresponding baseline distribution (unknown), and β is the regression coefficient (possibly a vector, but for the simplicity of notation we only assume it is only one dimensional throughout) in which we are most interested. Hence we can write down the log-likelihood function in terms of F_0 and β. Application of the GP method is straightforward. Specifically, let

$(X_1, T_1, U_1, V_1, Z_1)$, ...,$(X_n, T_n, U_n, V_n, Z_n)$ be a sample of n (incomplete) observations, where X_i is the interested event time but not observed, (T_i, U_i, V_i) are the (left) truncation and (interval) censoring times observed, and Z_i is the covariate observed. As before, let τ_i, $i = 1, ..., k$, be the order statistics of T_i, U_i and V_i, $1 \le i \le n$, which may be further reduced by Turnbull's algorithm. With covariates, it is more convenient to parametrize the baseline distribution via its cumulative hazard function $H(.)$. The log-likelihood divided by n can be written as:

$$\psi(H_0, \beta) := \int_{R^5} \phi_{(H_0,\beta)}(x, t, u, v, z) dP_n(x, t, u, v, z),$$

where P_n is the empirical probability measure of the $(X_i, T_i, U_i, V_i, Z_i)$, $1 \le i \le n$, and

$$\begin{aligned}\phi_{(H_0,\beta)}(x, t, u, v, z) := {} & I(t \le x \le u) \log\{\exp[-H_0(t)\exp(z\beta)] - \exp[-H_0(u)\exp(z\beta)]\} \\ + {} & I(u < x \le v) \log\{\exp[-H_0(u)\exp(z\beta)] - \exp[-H_0(v)\exp(z\beta)]\} \\ - {} & I(x > v) H_0(v)\exp(z\beta) + H_0(t)\exp(z\beta).\end{aligned}$$

We need to modify the definition of the process $W(t)$: for $t \ge 0$,

$$\begin{aligned}W_{(H_0,\beta)}(t) = {} & \int_{y\in[0,t], t'\le x\le y} \frac{\exp[-H_0(y)\exp(z\beta)]\exp(z\beta)}{\exp[-H_0(t')\exp(z\beta)] - \exp[-H_0(y)\exp(z\beta)]} dP_n(x, t', y, v, z) \\ - {} & \int_{y\in[0,t], y\le x\le u} \frac{\exp[-H_0(y)\exp(z\beta)]\exp(z\beta)}{\exp[-H_0(y)\exp(z\beta)] - \exp[-H_0(u)\exp(z\beta)]} dP_n(x, y, u, v, z) \\ + {} & \int_{y\in[0,t], u< x\le y} \frac{\exp[-H_0(y)\exp(z\beta)]\exp(z\beta)}{\exp[-H_0(u)\exp(z\beta)] - \exp[-H_0(y)\exp(z\beta)]} dP_n(x, t', u, y, z) \\ - {} & \int_{y\in[0,t], y< x\le v} \frac{\exp[-H_0(y)\exp(z\beta)]\exp(z\beta)}{\exp[-H_0(y)\exp(z\beta)] - \exp[-H_0(v)\exp(z\beta)]} dP_n(x, t', y, v, z) \\ - {} & \int_{y\in[0,t], x> y} \exp(z\beta) dP_n(x, t', u, y, z) + \int_{y\in[0,t]} \exp(z\beta) dP_n(x, y, u, v, z). \qquad (3)\end{aligned}$$

Denote $\Delta W_{(H,\beta),i} := W_{(H,\beta)}(\tau_i) - W_{(H,\beta)}(\tau_i^-)$, $i = 1, 2, ..., k$, $\Delta\beta_{(H,\beta)} := \partial\psi(H, \beta)/\partial\beta$. And let $\Delta W_{(H,\beta)} := (\Delta W_{(H,\beta),1}, ..., \Delta W_{(H,\beta),k})^T$. The corresponding gradient projection algorithm is the same as before:

$$\begin{aligned}dH_0^{(m+1)} &= \mathbf{Proj}(H_0^{(m)} + \Delta W_{(H_0^{(m)},\beta^{(m)})}) - H_0^{(m)}, \\ H_0^{(m+1)} &= H_0^{(m)} + \alpha_j dH_0^{(m+1)}, \quad \beta^{(m+1)} = \beta^{(m)} + \alpha_j \Delta\beta_{(H_0^{(m)},\beta^{(m)})}, \\ \alpha_j &= \max\left\{(1/2)^i : i \ge 0, \psi(H_0^{(m+1)}, \beta^{(m+1)}) > \psi(F_0^{(m)}, \beta^{(m)})\right\},\end{aligned}$$

where the projection is the same as above for the GP (without covariates). Again the projection step is taken to ensure that the estimate of F_0 is a proper distribution, which requires that H_0 is

Table 4: Mean and standard deviation (in parentheses) of estimated regression coefficient and convergence speed (iterations and CPU time in seconds) with left truncated and interval censored data, with 500 resplications.

β	$\hat{\beta}$	CPU time	Iterations
0.0	-0.006 (0.154)	64.8 (40.6)	655 (398)
0.2	0.188 (0.140)	39.1 (29.3)	637 (466)
0.5	0.445 (0.139)	34.8 (34.9)	405 (399)

nondecreasing. For some theoretical problems in the Cox proportional hazards model, such as the nonidentifiability with truncated data, the reader is referred to Alioum and Commenges (1996).

We note that it is also straightforward to extend the DICM to the Cox model (see Appendix). As in the case without covariates, the extended DICM works well for interval censored data but not for data also with truncation.

In our simulations, the baseline distribution F_0 is Weibull(2,1) with hazard function $h_0(x) = 2x$. There is only one covariate, $Z = 0$ or 1. Three values of β were varied: 0, 0.2 and 0.5. The random samples were generated in a way very similar to that in Section 3. The only difference is that we have two followups (instead of one in Section 3). Specifically $T_i \sim U(0,1)$ and $len = 0.5$. The sample size is around 200, in which almost half of them take covariate $Z = 0$ and another half $Z = 1$. The starting values for the baseline distribution and β were the discrete uniform and 0 respectively. The results are shown in Table 4. It seems the larger β, the shorter the time taken to converge. The slight under-estimation of the positive β is really not surprising, since the NPMLE tends to under-estimate the survival function $1 - F(x)$ (here the baseline $1 - F_0(x)$) for left truncated and interval censored data with only moderate sample sizes (Pan and Chappell, 1996). More accurate estimation procedures are currently under investigation.

5. Example

Now we apply the methodology to the Massachusetts Health Care Panel Study (MHCPS) (Chappell, 1991). The MHCPS was a statewide probability sample of noninstitutionalized people

age 65 and older living in the community. Special interest was paid to dependence upon others as indicated by the ability to fulfill measures of function called activities of daily living. A baseline survey was conducted in 1974, followed by a second examination fifteen months later. The third and fourth examinations were undertaken six and ten years respectively after the baseline. Only subjects who were functionally independent at baseline were included in the study, by which the left truncation was introduced. Interval censoring happened because we only know the event happened between two examinations, or right censored at the fourth one. We consider the non-poor male group and the non-poor female group with 421 and 609 subjects respectively. We are interested in the age when people lose their functional independence. The estimates of the survival (defined being functionally independent) probabilities computed by the EM, the DICM and the GP are shown in Figures 2 and 3, and their convergence speeds and the accuracies (including those from the DICM-EM and GP-EM) are provided in Table 5. We note that the EM took ver long but still *prematurely* stopped. This premature convergence resulted in a large difference in the shape of the survival curve. However, the difference is mainly caused by the first drop of the survival probability at age 65.3. If we condition on survival after age 65.3, the difference becomes minor. In this sense the EM is still stable. It is evident that the GP-EM also prematurely stopped as the DICM-EM, though it did very well as shown in Table 1 and Table 3. This confirms the result in Table 2 that the hybrid algorithm may also be less accurate. In this example, the DICM is still the fastest but less accurate than the GP.

To compare the distributions of two groups, the Cox proportional hazards model is adopted to do regression analysis. The non-poor female was taken as the baseline group, hence the female group with covariate value $Z = 0$ and the male with $Z = 1$. The estimated regression coefficient from the GP method is 0.16. It took 130.84 seconds of the CPU time with 782 iterations. It is fast enough to implement the bootstrap to conduct the statistical inference (Efron and Tibshirani, 1993). With 1000 bootstrap replications, the mean and standard deviation of $\hat{\beta}$ are 0.15 and 0.09 respectively. The bootstrap confidence intervals are shown in Table 6. The bootstrap percentile confidence interval and BC_a confidence interval were applied, which are first and second order accurate respectively. They are both close to the result of using the likelihood ratio test for $H_0 : \beta = 0$, which yields the p-value 0.088. Therefore it appears that there is some but not a highly significant difference between the survival (or distribution) functions of the two groups.

Table 5: Convergence speed (iterations and CPU time in seconds needed) and the maximized log-likelihood value at the convergence with MHCPS dataset.

	Non-poor female				
	EM	DICM	GP	DICM-EM	GP-EM
CPU time	6046.16	3.48	66.70	3.37	2.95
Iterations	38012	61	514	26	22
log-likelihood	-603.991	-598.578	-598.386	-610.250	-628.177

	Non-poor male				
	EM	DICM	GP	DICM-EM	GP-EM
CPU time	5700.09	3.13	29.10	1.61	1.62
Iterations	37839	84	350	18	18
log-likelihood	-440.143	-436.834	-436.221	-447.307	-464.098

Table 6: Bootstrap confidence intervals for regression coefficient with MHCPS dataset.

Confidence Level	Percentile C.I.	BC_a C.I.
85%	[0.031, 0.278]	[0.032, 0.280]
90%	[0.013, 0.296]	[0.013, 0.296]
95%	[-0.016, 0.320]	[-0.016, 0.320]
99%	[-0.061, 0.356]	[-0.061, 0.356]

6. Summary

Even though the EM algorithm is intuitively appealing, computationally it is not efficient in computing the NPMLE of distribution functions with interval censored data, especially when data also subject to truncation. In fact, it is so slow that practically it is not usable for large datasets. Many proposed methods of accelerating the EM by using the full Hessian or its approximation are also not applicable due to the large space required. Theoretically the ICM can be extended to the truncated data, but it seems more natural to implement it via a damped version. Moreover, by analyzing the ICM, we come to the gradient projection method which is fast and general enough to be widely applicable, such as in the Cox Proportional Hazards Model for (left-truncated and) interval censored data and in computing the monotone MLE (Pan and Chappell, 1996). With truncated (especially also moderately large) data, the EM, the DICM and the hybrids DICM-EM and GP-EM all may prematurely stop when the log-likelihood increment is taken as the only convergence criterion. However, the GP is stable and fast to achieve an accurate estimate. The premature convergence may be avoided by imposing more stringent convergence criteria, such as considering the difference of two consecutive estimates in addition to the log-likelihood increment. The unstable performance from the DICM with truncated data may also be caused by the crude approximation to the Hessian matrix of the log-likelihood. Further refinement along this direction may improve its performance.

Appendix. An extended damped ICM to the Cox model with left truncated and interval censored data

Unless specified otherwise, we use the same notation as in Section 4. Define the process G: for $t \geq 0$,

$$
\begin{aligned}
G_{(H_0,\beta)}(t) = & \int_{y\in[0,t],t'\leq x\leq y} \frac{\exp\{-[H_0(y)+H_0(t')]\exp(z\beta)\}\exp(2z\beta)}{\{\exp[-H_0(t')\exp(z\beta)]-\exp[-H_0(y)\exp(z\beta)]\}^2} dP_n(x,t',y,v,z) \\
+ & \int_{y\in[0,t],y\leq x\leq u} \frac{\exp\{-[H_0(y)+H_0(u)]\exp(z\beta)\}\exp(2z\beta)}{\{\exp[-H_0(y)\exp(z\beta)]-\exp[-H_0(u)\exp(z\beta)]\}^2} dP_n(x,y,u,v,z) \\
+ & \int_{y\in[0,t],u<x\leq y} \frac{\exp\{-[H_0(y)+H_0(u)]\exp(z\beta)\}\exp(2z\beta)}{\{\exp[-H_0(u)\exp(z\beta)]-\exp[-H_0(y)\exp(z\beta)]\}^2} dP_n(x,t',u,y,z) \\
+ & \int_{y\in[0,t],y<x\leq v} \frac{\exp\{-[H_0(y)+H_0(v)]\exp(z\beta)\}\exp(2z\beta)}{\{\exp[-H_0(y)\exp(z\beta)]-\exp[-H_0(v)\exp(z\beta)]\}^2} dP_n(x,t',y,v,z). \qquad (4)
\end{aligned}
$$

The process $W_{(H,\beta)}(t)$ is defined as in (3).

Denote $\Delta W_{(H,\beta),i} := W_{(H,\beta)}(\tau_i) - W_{(H,\beta)}(\tau_i^-)$, $\Delta G_{(H,\beta),i} := G_{(H,\beta)}(\tau_i) - G_{(H,\beta)}(\tau_i^-)$, $i = 1, 2, ..., k$, $\Delta\beta_{(H,\beta)} := \partial\psi(H,\beta)/\partial\beta$. And let

$$\Delta W_{(H,\beta)} := (\Delta W_{(H,\beta),1}, ..., \Delta W_{(H,\beta),k})^T$$

,

$$\Delta G_{(H,\beta)} := (\Delta G_{(H,\beta),1}, ..., \Delta G_{(H,\beta),k})^T$$

, and $\Delta\tilde{G}$ is similarly defined as in (1). Then our extended damped ICM algorithm is as follows:

$$H_0^{(m+1)} = \mathbf{Proj}(H_0^{(m)} + \alpha_j[\Delta\tilde{G}_{(H_0^{(m)},\beta^{(m)})}]^{-1}\Delta W_{(H_0^{(m)},\beta^{(m)})}),$$

and

$$\beta^{(m+1)} = \beta^{(m)} + \alpha_j\Delta\beta_{(H_0^{(m)},\beta^{(m)})},$$

where α_j is chosen to satisfy

$$\alpha_j = \max\left\{(1/2)^i : i \geq 0, \psi(H_0^{(m+1)}, \beta^{(m+1)}) > \psi(F_0^{(m)}, \beta^{(m)})\right\},$$

and the projection is defined by

$$\mathbf{Proj}(y) := \arg\min_x \{\sum_{i=1}^{k}(y_i - x_i)^2\Delta\tilde{G}_{x,i} : 0 \leq x_1 \leq x_2 \leq ... \leq x_k\}.$$

Acknowledgements

The first author would like to thank Yunlei Zhang and Xiwu Lin for their stimulating discussions. This research was supported by the Grant 1R29-EY10769-01 from the National Eye Institute.

References

Alioum, A. and D. Commenges, A Proportional Hazards Model for Arbitrarily Censored and Truncated Data. *Biometrics* **52**, (1996) 512-524.

Aragon, J. and D. Eberly, On Convergence of Convex Minorant Algorithms for Distribution Estimation with Interval-Censored data. *Journal of Computational and Graphical Statistics* **1**, (1992) 129-140.

Atkinson, S., The Performance of Standard and Hybrid EM Algorithms for ML Estimates of the Normal Mixture Model with Censoring. *Journal of Statistical Computation and Simulation* **44**, (1992) 105-115.

Bertsekas, D.P., On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control* **21**, 2, (1976) 174-183.

Branch, L.G. and L. Ku, Transition probabilities to dependency, institutionalization, and death among the elderly over a decade. *Journal of Aging and Health* **1**, (1989) 370-408.

Chappell, R., Sampling design of multiwave studies with an application to the Massachusetts Health Care Panel Study. *Statististics in Medicine* **10**, (1991) 1945-1958.

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

Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap*, (Chapman & Hall, New York, 1993).

Finkelstein, D.M., A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics* **42**, (1986) 845-854.

Frydman, H., A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *J. R. Stat. So c. Ser. B* **56**, (1994) 71-74.

Groeneboom, P. and J.A. Wellner, Information bounds and nonparamet ric maximum likelihood estimation. *DMV Seminar Band* **19**, (Birhauser Verlag, Basel, 1992).

Huang, J., Efficient estimation for the Cox Model with interval censoring. Tech. Rept. 274, (Dept. of Statistics, Univ. of Washington, Seattle, WA, 1994).

Huang, J. and J.A. Wellner, Efficient estimation for the Proportional Hazards Model with "Case 2" interval censoring. Tech. Rept., (Dept. of Statistics, Univ. of Washington, Seattle, WA, 1995).

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

Lange, K., A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of the Royal Statistical Society, Series B* **57**, (1995a) 425-437.

Lange, K., A Quasi-Newton Acceleration of the EM Algorithm. *Statistica Sinica* **5**, (1995b) 1-18.

Liu, C. and D.R. Rubin, The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence. *Biometrika* **81**, (1994) 633-648.

Meilijson, I., A Fast Improvement to the EM Algorithm on its own Terms. *Journal of the Royal Statistical Society, Series B* **51**, (1989) 127-138.

Meng, X.L. and D.B. Rubin, Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika* **80**, (1993) 267-278.

Pan, W. and R. Chappell, Estimating Survival Curves with Left Truncated and Interval Censored Data under Monotone Hazards. Tech. Rept. 109, (Dept. of Biostatistics, Univ. of Wisconsin, Madison, WI, 1996).

Polak, E., *Computational Methods in Optimization*, (Academic Press, New York, 1971).

Rai, S.N. and D.E. Matthews, Improving the EM Algorithm. *Biometrics* **49**, (1993) 587-591.

Redner, R.A. and H.F. Walker, Mixture Densities, Maximum Likelihood, and the EM Algorithm. *SIAM Review* **26**, (1984) 195-239.

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

Satten, G., Rank-Based Inference in the Proportional Hazards Model for Interval Censored Data. *Biometrika* **83**, (1996) 355-370.

Sinha, D., M.A. Tanner and W.J. Hall, Maximization of the Marginal Likelihood of Grouped Survival Data. *Biometrika* **81**, (1994) 53-60.

Turnbull, B.W., The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38**, (1976) 290-295.

Zhan, Y., and J.A. Wellner, Double Censoring: Characterization and Computation of the Nonparametric Maximum Likelihood Estimator. *Tech. Rept.* (Dept. of Statistics, Univ. of Washington, Seattle, WA, 1995).

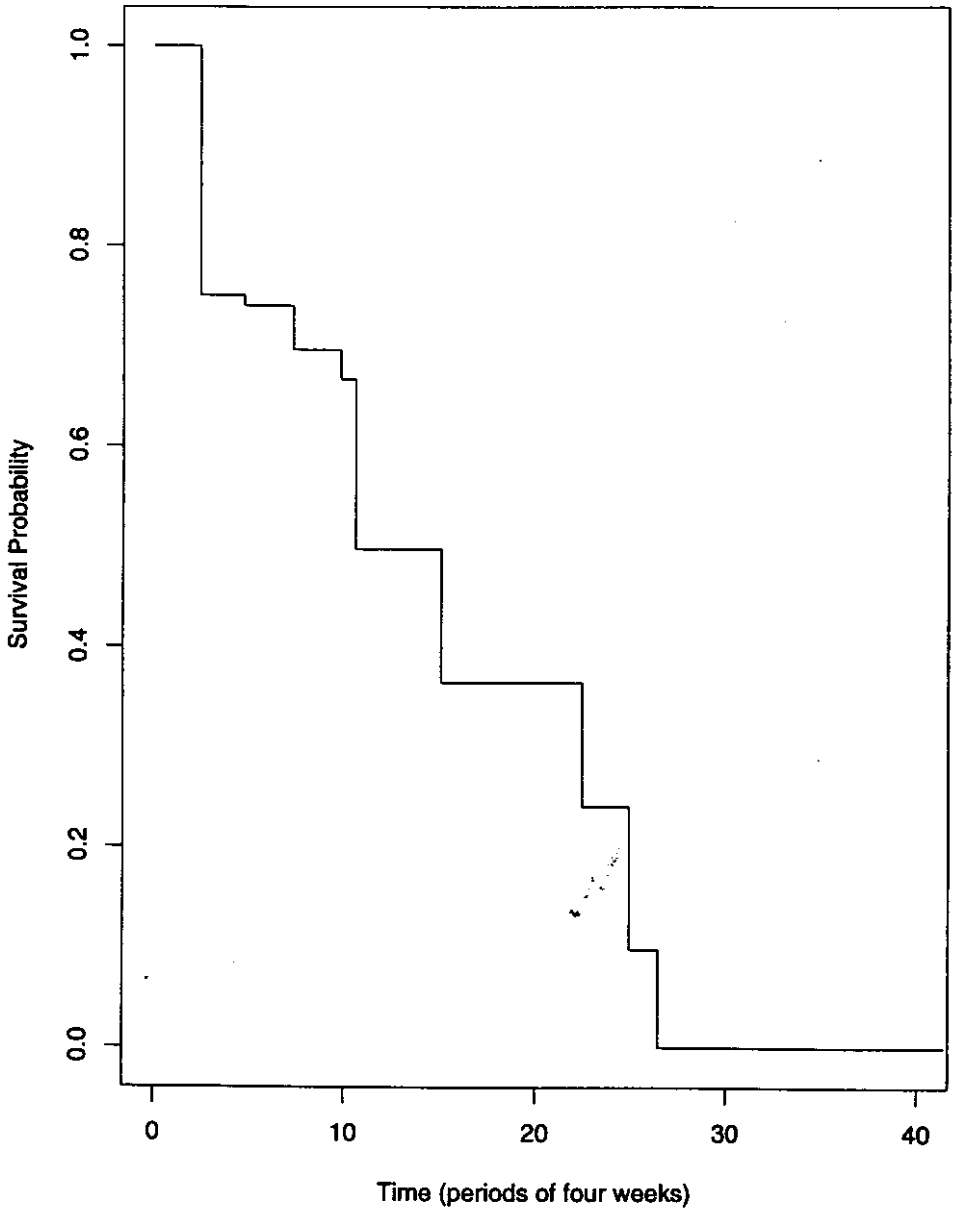
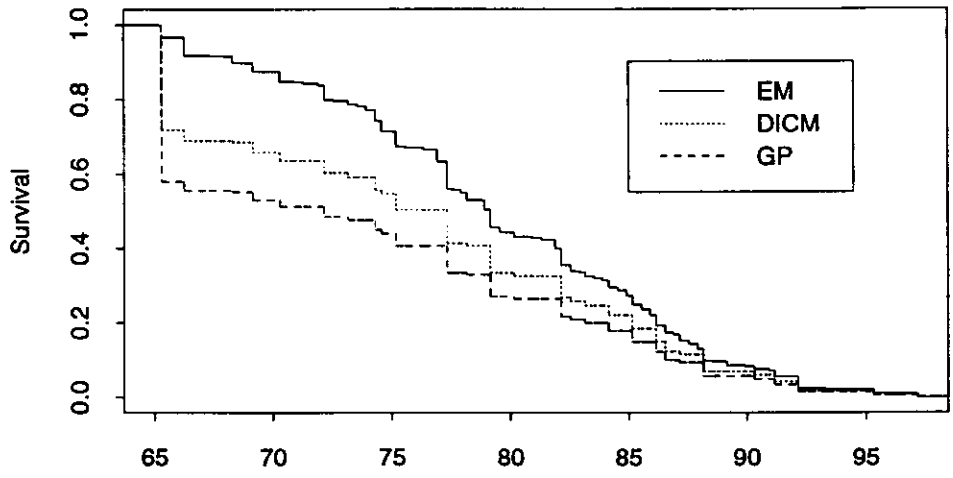


Figure 1: *Estimates of the survival probability from the AIDS data.*



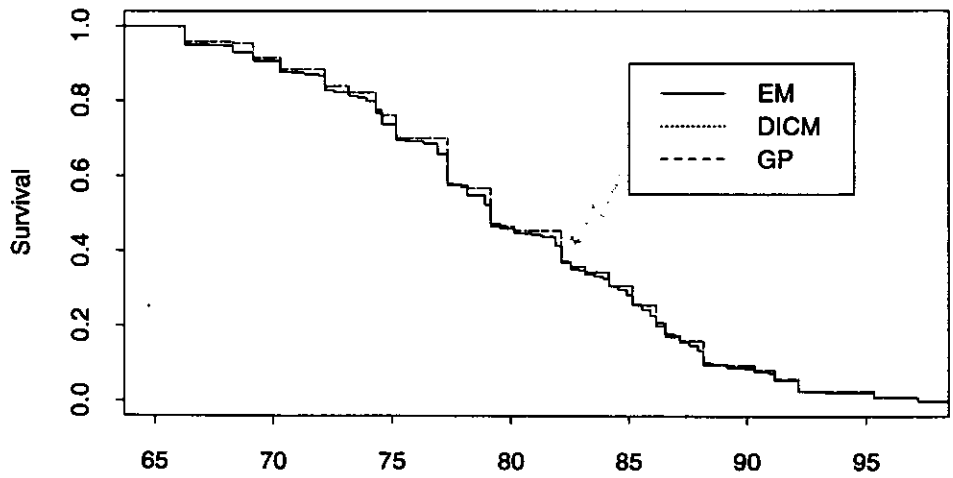


Figure 2: *Estimates of the survival probability for the non-poor female group from the MHCPS.*

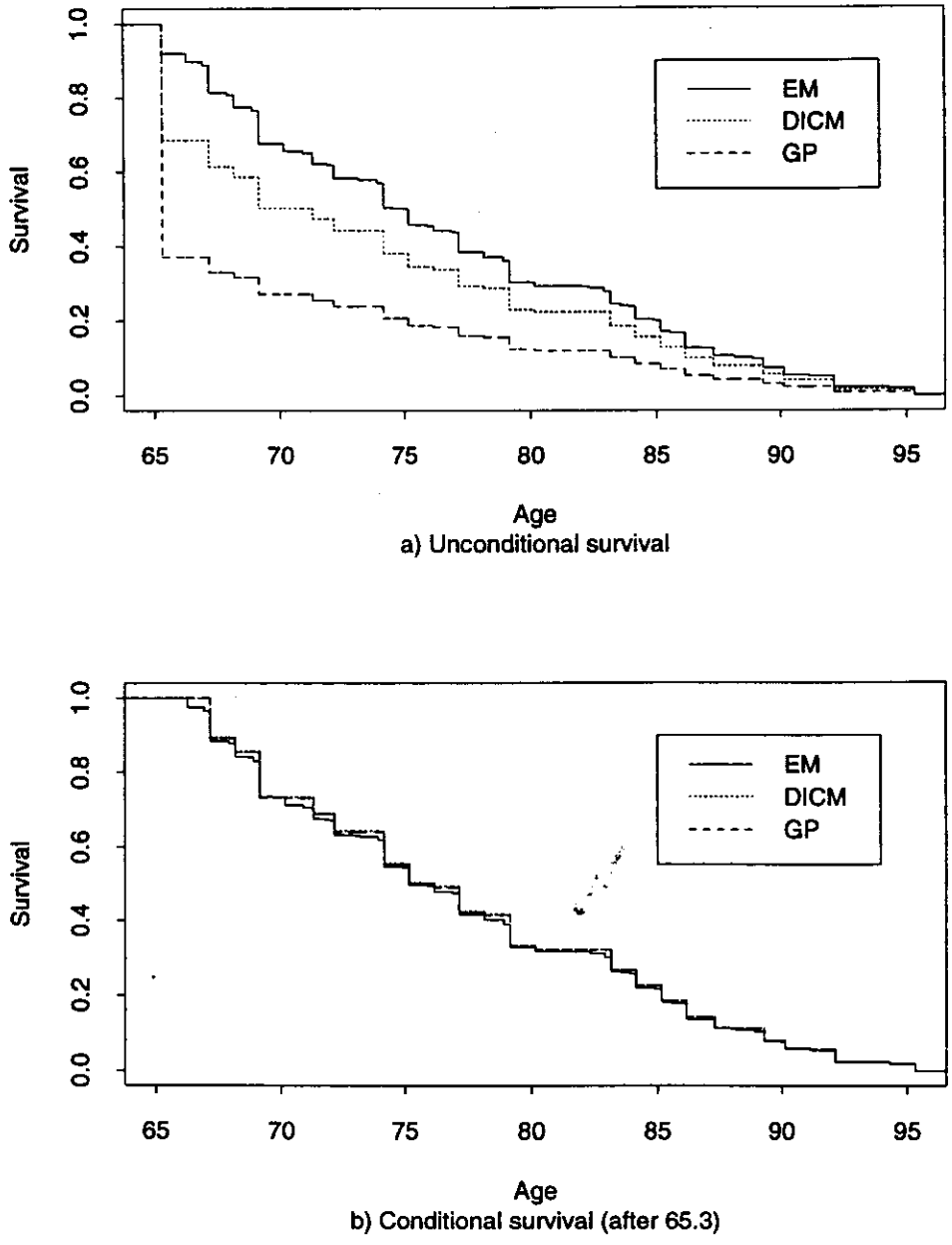


Figure 3: *Estimates of the survival probability for the non-poor male group from the MHCPS.*