UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS

Technical Report #118

February 1997

Estimating survival curves with left truncated and interval censored data via the EMS algorithm

Wei Pan
Richard Chappell, Ph.D.

Department of Biostatistics

MADISON, WISCONSIN

ESTIMATING SURVIVAL CURVES WITH LEFT TRUNCATED AND INTERVAL CENSORED DATA VIA THE EMS ALGORITHM

Wei Pan and Rick Chappell

Departments of Biostatistics and Statistics, University of Wisconsin, K6/430 CSC, 600 Highland Ave., Madison, Wisconsin 53792, U.S.A.

Keywords and Phrases: EM algorithm; maximum penalized likelihood; nonparametric maximum likelihood.

ABSTRACT

It is well-known that the nonparametric maximum likelihood estimator (NPMLE) of a survival function may *severely* under-estimate the survival probabilities at very early times for left truncated data. This problem might be overcome by instead computing a smoothed nonparametric estimator (SNE) via the EMS algorithm. The close connection between the SNE and the maximum penalized likelihood estimator is also established. Extensive Monte Carlo simulations show superior performance of the SNE over that of the NPMLE, in terms of either bias or variance, even for moderately large samples. The methodology is illustrated with the application to the Massachusetts Health Care Panel Study dataset to estimate the probability of being functionally independent for non-poor male and female groups respectively.

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. In addition to censoring, truncation may also arise, such as in the Massachusetts Health Care Panel Study (Chappell, 1991a). In these studies a baseline examination is first conducted, then after some fixed time the first followup examination takes place, then second and third, *etc.* followups (if any) after intervals of several years each. 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. Left truncation is introduced because we condition on the event of interest (such as an infection of a disease) happening after the entries of the subjects. Turnbull (1976) gives an EM algorithm (Dempster, Laird and Rubin, 1977) to compute the nonparametric maximum likelihood estimate (NPMLE) of a survival function $S(.)$ for arbitrarily truncated and censored data. Frydman (1994) points out an error in Turnbull's characterization of the support of the NPMLE when applied to truncated data. (Throughout this paper, Turnbull's algorithm is corrected as suggested by Frydman.) Though the NPMLE may have some desirable large sample properties, such as consistency (Tsai, Jewell and Wang, 1987, for left-truncated and right-censored data), it is well-known that with a finite sample it may *severely* under-estimate the survival probabilities at very early times for left truncated data (Tsai, 1988, for left truncated and right censored data; Pan and Chappell, 1996a, for left truncated and interval censored data). One efficient approach is through the monotone MLE, i.e. the MLE under a nondecreasing hazard assumption (Tsai, 1988, for left truncated and right censored data; Pan and

Chappell, 1996a, for left truncated and interval censored data). But it seems that there is still no well established method to check the monotone hazard assumption for truncated and censored data. Chappell (1991b) gives a "redistributed NPMLE" with some primitive smoothing idea. Pan and Chappell (1996b) present a nonparametric estimator based on the Nelson-Aalen estimator (Nelson, 1969), which alleviates the under-estimation problem from the NPMLE and may also be used as a diagnostic tool for the monotone MLE. In this paper, a smoothed nonparametric estimator (SNE) is proposed and it can be easily computed through the EMS algorithm (Silverman, Jones, Wilson and Nychka, 1990).

The EMS algorithm is a simple modification of the well-known EM algorithm, which adds a smoothing (S) step after the usual expectation (E) and maximization (M) steps. It has already been shown that the EMS algorithm may yield better estimates than the EM algorithm in some applications such as image analysis and integral equations. Nychka (1990) presents some theoretical properties of the EMS algorithm including its close connection with the maximum penalized likelihood method. Specifically, for a density estimation model, "it is shown that the limit point of the EMS algorithm (using a particular weighted smoother) is also an extremum of a penalized likelihood. In fact, each EMS iteration is very similar to the EM step for maximizing this penalized likelihood." And "This identification not only makes the EMS algorithm less *ad hoc* but also suggests why the algorithm converges." Leung and Elashoff (1996) apply the EMS algorithm to estimate the underlying distributions in a three-state disease model with interval censored data. An apparent possible advantage of using the EMS algorithm is to yield a smoothed estimate (where the "smooth" is in a more general sense that jump sizes of the estimate does not varying as rapidly and the estimates still is not smooth in the usual sense). But our main motivation here

is to use it to overcome the under-estimation problem caused by the NPMLE.

The remainder of the paper is organized as follows. First the EMS algorithm (and the EM algorithm) and the SNE are briefly introduced. Section 3 establishes the connection between the SNE and the maximum penalized likelihood estimator. In Section 4 some simulation results are presented to show the superior performance of the SNE over the NPMLE. At last our method is illustrated by application to the MHCPS and Channing House datasets.

2. THE EMS ALGORITHM AND THE SNE

The EMS algorithm proceeds by iterating three steps: E-, M- and S-steps until convergence (by some criterion). The E- and M-steps are the same as in Turnbull (1976), both of which have simple closed solutions. For completeness, in addition to S step we also briefly review the E- and M-steps in Turnbull's algorithm. Turnbull has shown that the NPMLE can only have probability masses in a finite number of disjoint intervals denoted as $[a_i, b_i], i = 1, ..., n$. These intervals are defined by the data but are not in general the intervals of censorship. Suppose the true conditional probability in each interval is p_i and the corresponding estimate at the kth iteration is $\hat{p}_i^{(k)}$. Denote the number of events of interest happening in $[a_i, b_i]$ as D_i. Then the E-step is simply to compute $d_i^{(k)} = E(D_i|\{\hat{p}_i^{(k)}\}, data)$. The M-step is accomplished by obtaining a new estimate:

$$\hat{p}_i^{(k+1)} = d_i^{(k)}/M^{(k)}, \qquad M^{(k)} = \sum_{j=1}^{n} d_j^{(k)}.$$

Turnbull's EM algorithm proceeds by repeating the above two steps until convergence, which results in the NPMLE. For the EMS algorithm, another S-step is

added in *each* iteration. The S-step is often realized as a weighted average of some neighboring estimates. In our case, it is simply:

$$\hat{p}_i^{(k+1)} = \frac{1}{2^{2j}} \sum_{b=-j}^{j} \begin{pmatrix} 2j \\ b+j \end{pmatrix} \tilde{p}_{i+b}^{(k+1)},$$

Such $(2j+1)$-point smoothing is usually applied with j values of 1,2,3 or 4. The larger is j, the more smoothing is performed. As suggested by Leung and Elashoff (1996), using $j = 1$ (corresponding to a triangular kernel) generally provides good estimates without incurring too much computation. This value is what we adopt throughout. When the above smoothing is applied to the two endpoints, some modifications are required to ensure the smoothed estimate constitutes a proper distribution function:

$$\hat{p}_1^{(k)} = (3\tilde{p}_1^{(k)} + \tilde{p}_2^{(k)})/4, \qquad \hat{p}_{n_s}^{(k)} = (3\tilde{p}_{n_s}^{(k)} + \hat{p}_{n_s-1}^{(k)})/4,$$

where $n_s = n - 1$ if $b_n = \infty$; or $n_s = n$ otherwise. Since if the last interval is infinite (dut to right-censoring), it does not make sense to smooth it or use it to smooth its neighbor.

The above three steps are repeated until $\{\hat{p}_i^{(k)}\}$ converges. The resulting estimate at convergence is the SNE. It is worth to note that the effect of smoothing at each iteration in the EMS algorithm is much different from smoothing the NPMLE at the convergence of the EM algorithm.

3. THE EMS AND PENALIZED LIKELIHOOD

The smoothing step in the EMS algorithm naturally reminds one the penalized likelihood method. In fact, the close relation between the EMS algorithm and the maximum penalized likelihood estimate has clearly been established by Nychka (1990)

for a Poisson model of density estimation. The result also holds in our current situation. Namely, the SNE from the EMS algorithm is also an extremum of a penalized likelihood. We will try to use the same notation as in Trunbull (1976).

Generally we use $x = (x_1, ..., x_n)^T$ to denote a column vector; and any arithmetic operation on vectors is component-wise. Let the log-likelihood be $L(p)$, and $d(p) = \partial L/\partial p$ (Turnbull's expressions (3.6) and (3.8)). As in Nychka (1990), surprisingly the penalty term will depend on the square root of p instead of p itself. Define $\theta = \sqrt{p}$ and let the penalized likelihood function be

$$\ell(p; R) = L(p) - \frac{1}{2}\theta^T R\theta,$$

where symmetric R is the roughness penalty matrix. To maximize $\ell(p; R)$, we only need to solve the equations:

$$0 = \partial\ell(p; R)/\partial p = 2D(p)\theta - 2R\theta \tag{1}$$

where $D(p) = diag(d_j(p))$.

On the other hand, by Turnbull's (3.10), at each iteration the EM algorithm outputs

$$\tilde{p}^{(k+1)} = \left\{I + D(\hat{p}^{(k)})/M^{(k)}\right\}\hat{p}^{(k)},$$

and the S-step in the EMS algorithm in a general form is: $\hat{p}^{(k+1)} = S_k\tilde{p}^{(k+1)}$. Here the smoothing matrix is $S_k = \Theta_k U \Theta_k^{-1}$, where $\Theta_k = diag(\theta_j)$, and U is a symmetric and full rank matrix.

At convergence,

$$\tilde{p}^* = \{I + D(\hat{p}^*)/M^*\}\hat{p}^*, \qquad \hat{p}^* = S^*\tilde{p}^*.$$

Hence for any $j = 1, ..., n$,

$$0 = d_j(\hat{p}^*)\theta_j^* - \frac{M^*}{\theta_j^*}\left[\tilde{p}_j^* - \hat{p}_j^*\right]. \tag{2}$$

But

$$\tilde{p}^* - \hat{p}^* = \Theta^*(U^{-1} - I)\Theta^{*-1}\hat{p}^*$$

and thus

$$\frac{1}{\theta_j^*}\left[\tilde{p}_j^* - \hat{p}_j^*\right] = \left[(U^{-1} - I)\theta^*\right]_j .$$

Substituting the last expression into (2), we obtain the same equation as (1) **if** $R = M^*(U^{-1} - I)$.

Therefore, the resulting SNE $\hat{p}^*$ is also an extreme point of the above penalized likelihood. Notice that the penalty term depends on $\hat{p}^*$ through M^* (as a "smoothing parameter"). Since U is the central part of the smoothing matrix, it is intuitive to think U^{-1} (and hence R) as a roughness penalty matrix.

4. MONTE CARLO SIMULATIONS

In this section, we will show selected results of our Monte Carlo simulations of the SNE. Suppose X is the survival time of the event of interest (hence we are estimating $S(X)$) and is independent of T, the baseline examination time. Y is the time of the first followup and m is the number of subjects. The time between the baseline survey and the first followup is $len = Y - T$, which is taken to be constant to mimic the pattern of the MHCPS data and other panel surveys. Only one followup takes place, which results in a high percentage of right-censoring, making the problem more challenging. 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 from 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 interval-censored in $[t_i, t_i + len)$; else it is right-censored at $t_i + len$.

To facilitate comparison, results from the NPMLE computed via the EM algorithm are also presented here. The initial values for both the EM and EMS algorithms are taken consistently to be: $\hat{p}_1^{(0)} = ... = \hat{p}_n^{(0)} = 1/n$. And the convergence criterion is that $\max_{1 \leq i \leq n} |\hat{p}_i^{(k+1)} - \hat{p}_i^{(k)}| < 0.0001$. Notice that for both the NPMLE and the SNE, the specific probability distribution in each small interval is nonidentifiable. For comparison, in all of our simulations, we display them as if their probability is linear in each small interval. Because the intervals are short, this approximation should be adequate.

In Figures 1a) and b), the true distribution (of X) is standard exponential, $T \sim U(0, 1.5)$ (i.e. uniform in $[0, 1.5]$), $len = 0.5$, and sample size (that is random) is around 110, half of which are right censored. As a comparison, The NPMLE has much more *severe* under-estimation.

In Figures 1c) and d), the true distribution is Weibull with shape parameter 4 and scale parameter 1. $T \sim U(0, 1.5)$, $len = 0.5$, sample size is around 130 and half of them are right censored. The performance of the SNE seems very satisfactory in terms of bias and variation. But the NPMLE still under-estimates survival at early times.

In Figure 2, the true distribution is Gamma with shape parameter 2 and scale parameter 1, $T \sim U(0, 6)$, $len = 0.5$, and almost 80 percent of observations are right censored. To investigate the large sample properties, we increase sample size from near 100 to 1000. For both estimators, as their sample size increases both their bias and variance decrease. When sample size is around 1000, the SNE is very close to the true distribution. But there is still some evident under-estimation from the NPMLE

at early times, which is carried over at later times as well.

5. TWO REAL EXAMPLES

Now we apply the methodology to the Massachusetts Health Care Panel Study (MHCPS) (Chappell, 1991a). 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 was right censored at the fourth one. We consider the non-poor male and female groups with 421 and 609 subjects respectively. We are interested in the age when people lose their functional independence, the ability to live without full-time assistance. The SNE and the NPMLE of the "survival" (defined as being functionally independent) probabilities are shown in Figure 3. It is not surprising that there is a big drop of each NPMLE at age 65.3, which seems unrealistic. If we condition on survival after age 65.3, the resulting conditional NPMLEs are very close to the (unconditional) SNEs (each of which is not much different from its own conditional obviously). Therefore we tend to believe that each (unconditional) NPMLE likely under-estimates its corresponding true survival probability.

Though all our discussion so far is concentrated on left truncated and interval censored data, it is easy to extend the SNE to arbitrarily truncated and censored data

due to the generality of the EMS algorithm. As an example, we apply the method to the well-studied Channing House data that is left truncated and right censored (Hyde, 1977). This dataset provides survival times in months for men in a Palo Alto retirement community. The truncation time is the age of the subject at entry into the community. Among 97 men, 46 died while 46 survived to the end of the study and 5 withdrew from the community (hence 51 were right-censored). As pointed out by Tsai (1988), there are two sharp drops from the NPMLE at early times (777 and 781 months) while the monotone MLE overcomes this under-estimation. From Figure 4, we can see the SNE also overcomes the problem well, though it may still be slightly under-biased at the beginning (when compared with the monotone MLE). Because of the closeness between the SNE and the monotone MLE, we confirm that the intuitive assumption of a nondecreasing hazard function is very likely to be appropriate here.

6. DISCUSSION

We have shown that the SNE from the EMS algorithm can overcome the under-estimation problem generated by the NPMLE for left truncated and interval censored data. One immediate application of the SNE is as a diagnostic tool for checking the monotone hazard assumption in the monotone MLE. Moreover, it is well known that the EM algorithm (in computing the NPMLE) may be very slow. Our experience confirms Leung and Elashoff's (1996), that the EMS may converge much faster than the EM algorithm. We also note that the SNE can be easily extended to arbitrarily truncated and censored data.

Though the smoothing scheme used in our EMS algorithm is simple, its performance is still satisfactory. But in some other applications, it may be better to explore some more complex smoothing schemes. It is also worthwhile to remind the reader

the limitations of many smoothing methods, such as the possible bias when the to-be estimated function is not "smooth". Another example is that if the survival function is discrete, our proposed SNE is more likely to be inconsistent. But some modifications, such as only smoothing when the risk set size is relatively small (i.e. less than some preset constant), will possibly improve its performance.

The current practice in estimating the regression coefficient in the Cox proportional hazards model with left truncated and interval censored data is to maximize the joint likelihood of the regression coefficient and the baseline survival (Alioum and Commenges, 1996). From our above discussion, it is expected that the joint NPMLE may under-estimate the baseline survival (Pan and Chappell, 1997). It is interesting to investigate whether our current work can be extended to this regression setting.

ACKNOWLEDGEMENTS

This research was supported by the National Eye Institute.

BIBLIOGRAPHY

Alioum, A. and D. Commenges, A proportional hazards model for arbitrarily censored and truncated data. *Biometrics* **52**, (1996) 512-524.

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

Chappell, R. (1991b). *Collection and analysis of truncated censored data.* University of Chicago Department of Statistics Ph.D. Thesis.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1-38.

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

Hyde, J. (1977). Testing survival under right censoring and left truncation. *Biometrika* vol. 64, 225-230.

Leung, K.-M. and Elashoff, R.B. (1996). A three-state disease model with interval-censored data: estimation and applications to AIDS and cancer. *Lifetime Data Analysis* **2**, 175-194.

Nelson, W.B. (1969). Hazard plotting for incomplete failure data. *J. Quality Technology* vol. 1, 27-52.

Nychka, N. (1990). Some properties of adding a smoothing step to the EM algorithm. *Statist. Probab. Lett.* **9**, 187-193.

Pan, W. and Chappell, R. (1996a). Estimating survival curves with left truncated and interval censored data under monotone hazards. Tech. Rept. 109, Biostatistics Dept., Univ. of Wisconsin-Madison.

Pan, W. and Chappell, R. (1996b). A nonparametric estimator of survival functions for arbitrarily truncated and censored Data. Tech. Rept. 111, Biostatistics Dept., Univ. of Wisconsin-Madison.

Pan, W. and Chappell, R. (1997). Estimation in the Cox Proportional Hazards Model with left truncated and interval censored data. Tech. Rept., Statistics Dept., Univ. of Wisconsin-Madison.

Silverman, B.W., Jones. M.C., Wilson, J.D. and Nychka, D.W. (1990). A smoothed EM approach to indirect estimation problems with particular reference to stereology and emission tomography (with discussion). *J. Roy. Statist. Soc.* B **52**, 271-324.

Tsai, W.Y. (1988). Estimation of the survival function with increasing failure rate based on left truncated and right censored data. *Biometrika* **75**, 2, 319-324.

Tsai, W.Y., Jewell, N.P. and Wang, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74**, 4, 883-886.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc.* B **38**, 290-295.

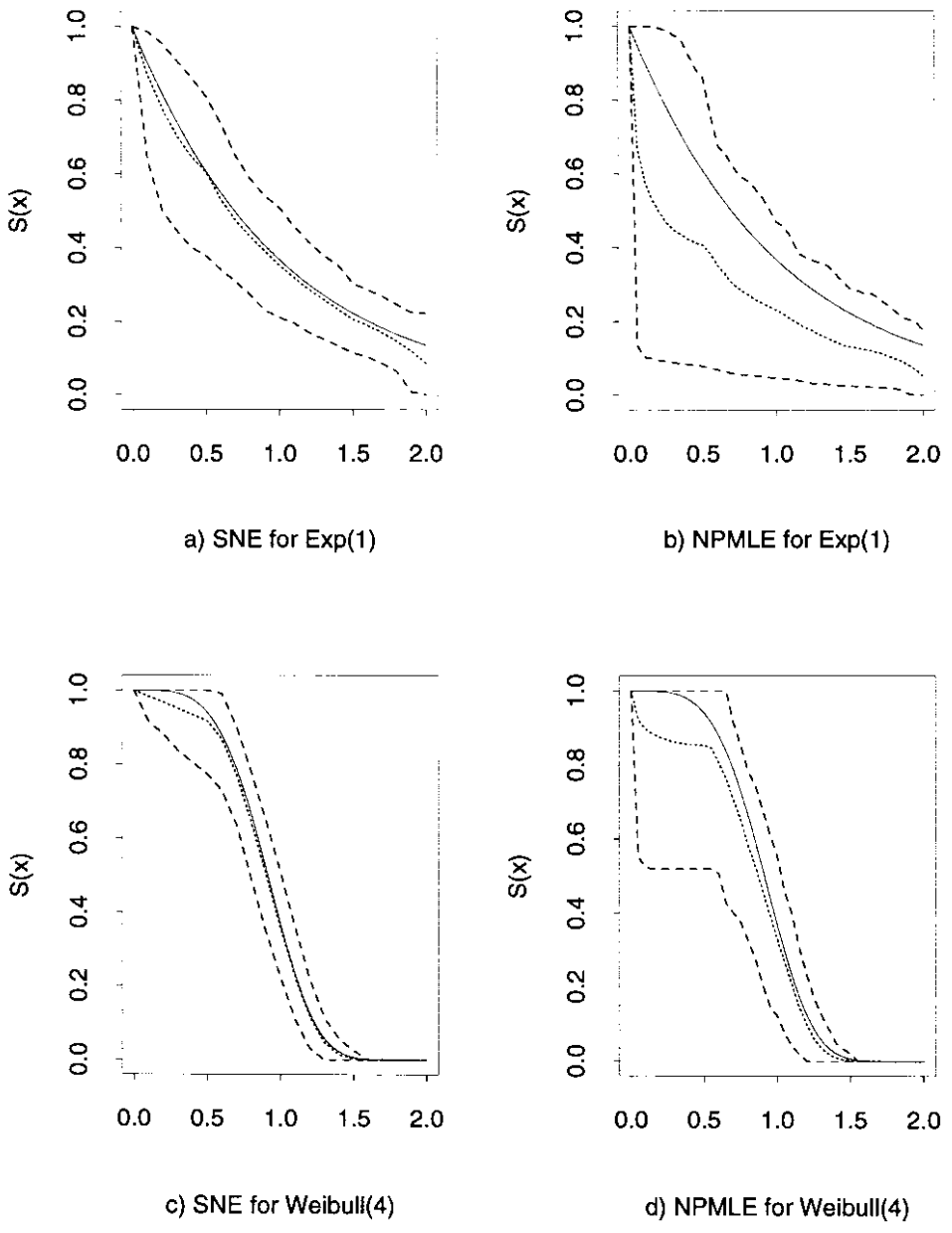

Figure 1: *Simulation results from* 500 *samples. The solid lines are true distributions, and the dotted ones are the means, the lower and upper* 5% *quantiles of the estimates. The maximum Monte Carlo standard errors are 0.0065, 0.0151, 0.0040 and 0.0086 respectively.*

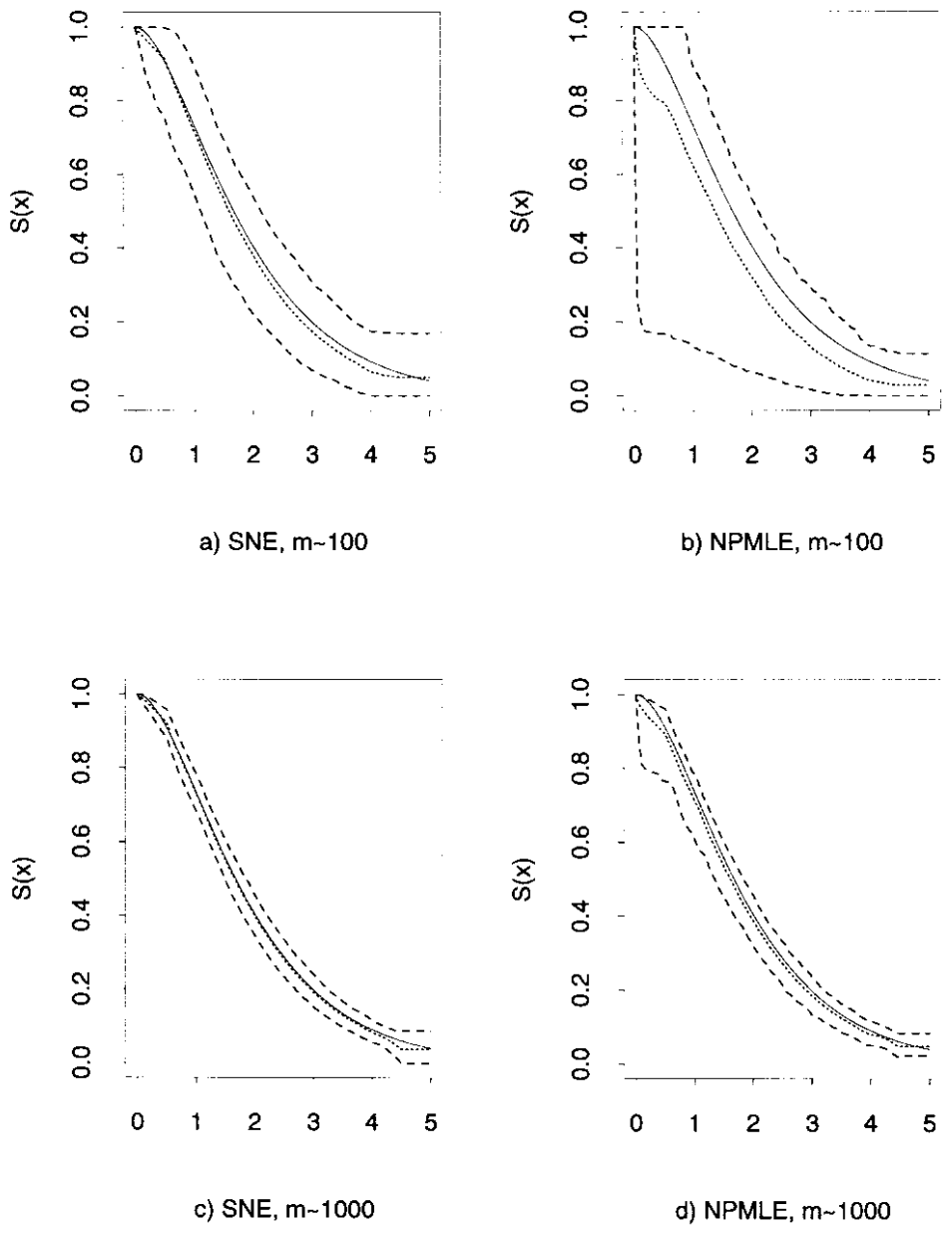

Figure 2: *Simulation results for Gamma*$(2,1)$ *with small to large sample sizes. The solid lines are true distributions, and the dotted ones are the means, the lower and upper* 5% *quantiles of the estimates from* 500 *samples. The maximum Monte Carlo standard errors are 0.0050, 0.0111 , 0.0015 and 0.0032 respectively.*

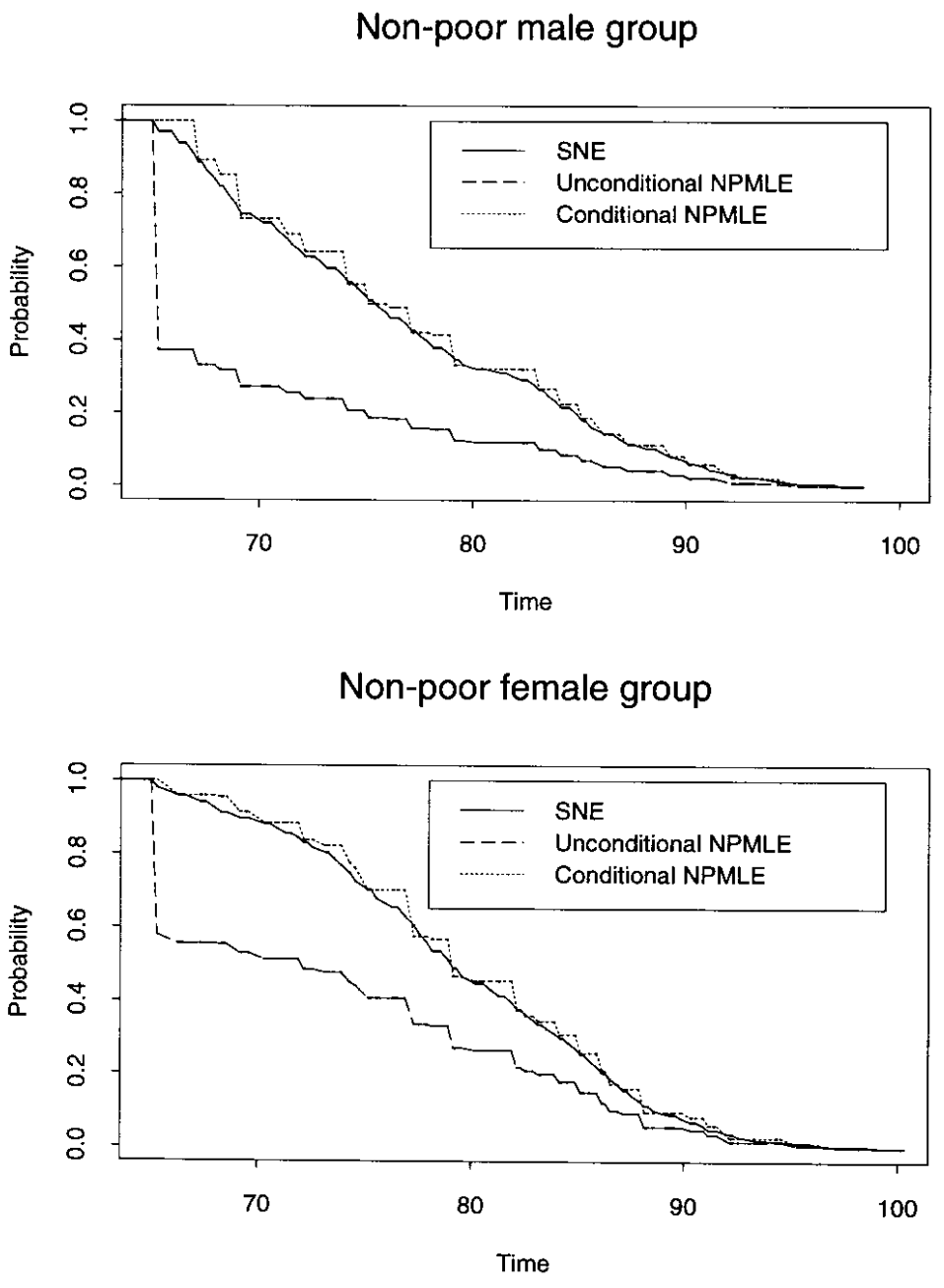

Figure 3: *Estimates of the survival probabilities for the non-poor male and female groups from the MHCPS.*

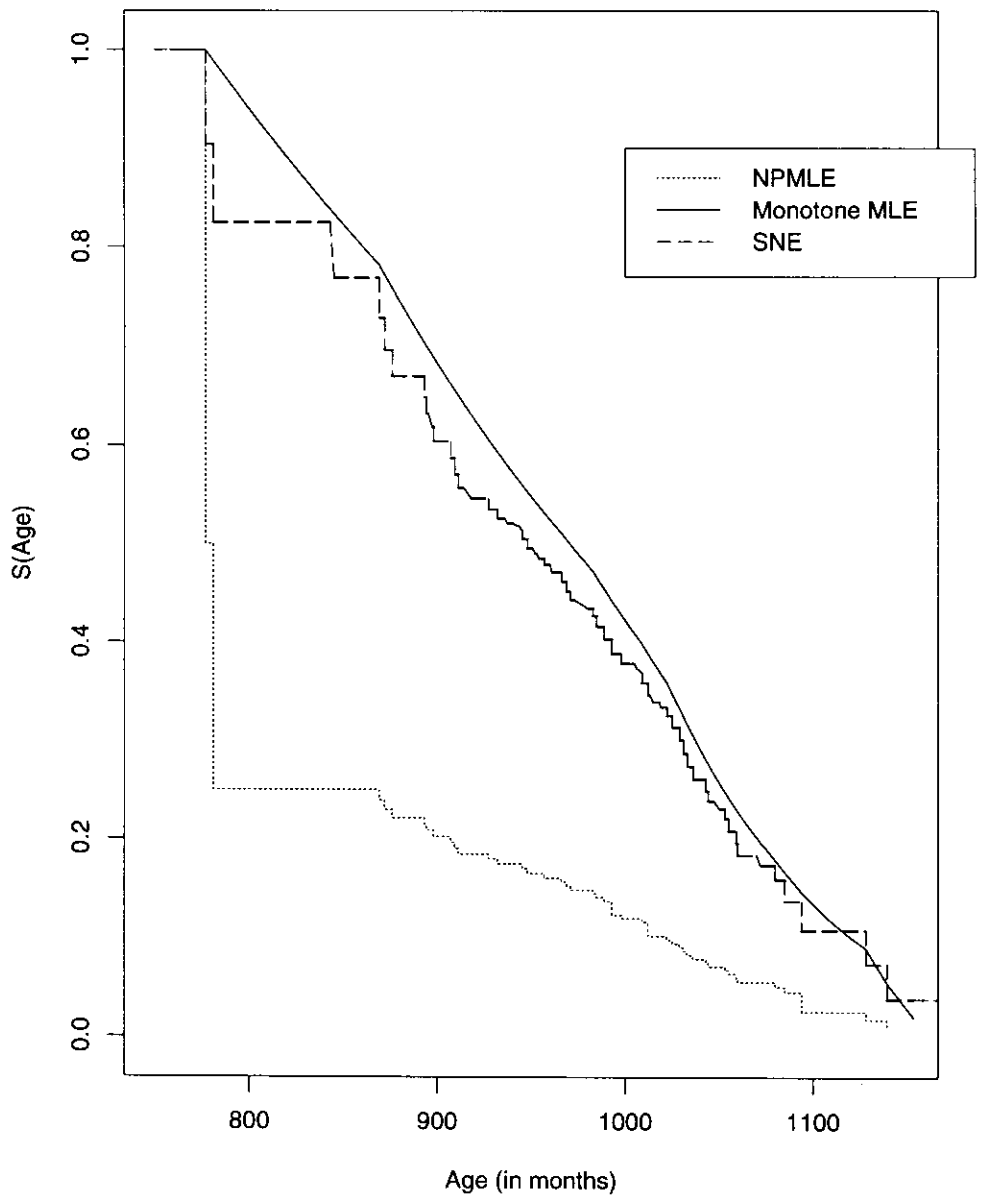

Figure 4: *Estimates of the (conditional) survival probability from the Channing House data.*