Lifetime Data Anal (2016) 22:473–503
DOI 10.1007/s10985-015-9345-9

Nonparametric estimation of time-to-event distribution based on recall data in observational studies

Sedigheh Mirzaei Salehabadi[1] · **Debasis Sengupta**[1]

Received: 14 October 2014 / Accepted: 20 July 2015 / Published online: 21 September 2015

Abstract In a cross-sectional observational study, time-to-event distribution can be estimated from data on current status or from recalled data on the time of occurrence. In either case, one can treat the data as having been interval censored, and use the nonparametric maximum likelihood estimator proposed by Turnbull (J R Stat Soc Ser B 38:290–295, 1976). However, the chance of recall may depend on the time span between the occurrence of the event and the time of interview. In such a case, the underlying censoring would be informative, rendering the Turnbull estimator inappropriate. In this article, we provide a nonparametric maximum likelihood estimator of the distribution of interest, by using a model adapted to the special nature of the data at hand. We also provide a computationally simple approximation of this estimator, and establish the consistency of both the original and the approximate versions, under mild conditions. Monte Carlo simulations indicate that the proposed estimators have smaller bias than the Turnbull estimator based on incomplete recall data, smaller variance than the Turnbull estimator based on current status data, and smaller mean squared error than both of them. The method is applied to menarcheal data from a recent Anthropometric study of adolescent and young adult females in Kolkata, India.

Keywords Informative censoring · Interval censoring · Nonparametric maximum likelihood estimator · Self consistency algorithm · Turnbull estimator

✉ Sedigheh Mirzaei Salehabadi
sedigheh_r@isical.ac.in

Debasis Sengupta
sdebasis@isical.ac.in

[1] Applied Statistical Unit, Indian Statistical Institute, Kolkata 700108, India

1 Introduction

Observational data on time to occurrence of a landmark event occur in various fields of biological and social sciences. Some examples of landmark events are onset of menarche in adolescent and young adult females (Bergsten-Brucefors 1976; Chumlea et al. 2003; Mirzaei et al. 2015), dental development (Demirjian et al. 1973; Eveleth and Tanner 1990), breast development (Cameron 2002; Aksglaede et al. 2009), beginning of criminal career (Hosmer and Lemeshow 1999), marriage and birth of the first child (Allison 1982), end of a work career (LeClere 2005) and end of a strike (Hosmer and Lemeshow 1999). The probability distribution of the time till occurrence of the event is useful for comparing two populations, setting benchmarks for individuals, setting policy objectives and so on. Estimation of that distribution is therefore an important inferential issue. Ideally one would like to observe a number of individuals continuously or periodically until the occurrence of the landmark event (Korn et al. 1997; McKay et al. 1998). However, researchers often opt for cross-sectional studies in order to save time and cost.

Cross-sectional studies can produce dichotomous data on the current status of an individual (whether or not the landmark event has occurred till the day of observation). A binary data regression model such as probit or logistic model, with time as the covariate, is often used for estimating the probability distribution function (Hediger and Stine 1987; Ayatollahi et al. 2002). It is also possible to estimate the distribution nonparametrically, by regarding the current status data as either left or right censored observations. The nonparametric maximum likelihood estimator (NPMLE) proposed by Turnbull (1976) for interval censored data has occasionally been used in this set-up (Keiding et al. 1996).

In some cross-sectional studies, a subject is asked to recall the time of the landmark event, in case it has already taken place. Such retrospective data can be incomplete. In many cases (e.g., when the event has not happened or the subject cannot recall when it had happened) one can specify only a range for the requisite time. Thus, data arising from retrospective studies are also interval-censored. In this situation, it is tempting to use the likelihood for interval censored data, leading to a parametric MLE or the NPMLE obtained by Turnbull (1976). In fact, there are instances when the Turnbull estimator has been used for studying the distribution of age at reaching a developmental landmark (see, e.g., Aksglaede et al. 2009). However, the censoring mechanism in this set-up is likely to depend on the time-to-event, thereby making the censoring informative. This is because of the fact that memory generally fades with time. As an example, for two subjects interviewed at the same age, the one with later onset of menarche is more likely to remember the date. It may be recalled that the Turnbull estimator is not meant for informatively censored data, and can have large bias when the censoring is informative, as confirmed by simulations reported in this paper.

We propose in Sect. 2 a new approach for estimating the time-to-event distribution by using the recall information through an informative censoring model. Under this model, the time of observation is assumed to be independent of the time-to-event, and the recall probability is regarded as a function of the gap time between the event and the observation. In Sect. 3, we check the identifiability of the distribution of interest. In Sect. 4, we derive the NPMLE under the model, establish its existence and asymptotic uniqueness and provide a self-consistency algorithm for computing it. We also provide a computationally simpler alternative that is asymptotically equivalent to the NPMLE. In Sect. 5, we show that both the NPMLE and its approximation are consistent estimators of the identifiable part of the underlying distribution. Results of Monte Carlo simulations and an illustrative data analysis are reported in Sect. 6 and 7, respectively. The data analysis is based on a study of menarcheal age of adolescent and young adult females, undertaken by the Indian Statistical Institute, Kolkata, where the landmark event is the onset of menarche. The conditions chosen for simulations are also in line with this application. Some concluding remarks are provided in Sect. 8. Proofs of all the results are given in the Appendix.

2 Model and likelihood

Consider a set of subjects having time of the occurrence of landmark event $T_1, \ldots, T_n$, which are samples from a common distribution F with density f and support $[t_{min}, t_{max}]$. Let these subjects be interviewed at times $S_1, \ldots, S_n$, respectively, chosen from a finite set $\mathcal{S}$. Let, for $i = 1, \ldots, n$, δ_i be the indicator of $T_i \leq S_i$, i.e., the event having had occurred on or before the time of interview.

In the case of current status data, one only observes (S_i, δ_i), $(i = 1, 2, \ldots, n)$. The corresponding likelihood, conditional on the time of interview, is

$$\prod_{i=1}^{n} [F(S_i)]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}, \tag{1}$$

where $\bar{F} = 1 - F$.

In a retrospective study, the subject may remember the exact time of the event. Let ε_i be the indicator of recalling that time. Thus, for the ith subject, it can be said that $T_i \in A_i$, where the set A_i is defined as follows in three cases. For the sake of simplicity, and in order to reduce the effect of recall error, we assume that whenever a subject forgets the epoch, there is no reliable information on an approximate range of time also.

Case (i). When $\delta_i = 0$, i.e., the landmark event for the ith subject did not occur till the time of interview, we have $A_i = [S_i \vee t_{min}, t_{max}]$.

Case (ii). When $\delta_i = 1$ and $\varepsilon_i = 1$, i.e., the landmark event for the ith subject did occur and the subject can exactly recall the date, we have $A_i = \{T_i\}$.

Case (iii). When $\delta_i = 1$ and $\varepsilon_i = 0$, i.e., the landmark event for the ith subject did occur and the subject cannot recall the exact date, we have $A_i = [t_{min}, t_{max} \wedge S_i]$

Note that there can be a gray area between the choices of A_i in cases (ii) and (iii), when the subject can recall an approximate date or a range of dates of the landmark event. In view of the possibility of recall error, noted by several researchers (Rabe-Hesketh et al. 2001; Beckett et al. 2001) we ignore such incomplete information and treat these as cases of no recall.

If the underlying censoring mechanism is presumed to be noninformative, then the likelihood arising from the above retrospective data, conditional on the time of interview, is

$$\prod_{i=1}^{n}\left[\{F(S_i)\}^{1-\varepsilon_i}\{f(T_i)\}^{\varepsilon_i}\right]^{\delta_i}[\bar{F}(S_i)]^{1-\delta_i}, \tag{2}$$

However, it has been mentioned in Sect. 1 that the censoring mechanism is likely to be informative. Specifically, the non-recall probability, $P(\varepsilon_i = 0|\delta_i = 1)$ may depend on the time elapsed since the time of that event, $S_i - T_i$. We model the conditional probability of forgetting the date as an unspecified function of the elapsed time,

$$\pi(s-t) = P(\varepsilon_i = 0|T_i = t, S_i = s), \qquad s > t. \tag{3}$$

According to this model, the likelihood, conditional on the ages at interview, is

$$\prod_{i=1}^{n}\left[\left(\int_0^{S_i} f(u)\pi(S_i-u)du\right)^{1-\varepsilon_i}\{f(T_i)(1-\pi(S_i-T_i))\}^{\varepsilon_i}\right]^{\delta_i}[\bar{F}(S_i)]^{1-\delta_i}. \tag{4}$$

Here, the informativeness of the censoring mechanism is captured through the function π. When π is a constant, the likelihood (4) becomes a constant multiple of (2). As a further special case, if $\pi = 1$, then the likelihood (4) reduces to (1). On the other hand, when $\pi = 0$, i.e., the landmark event times are perfectly recalled, the product likelihood (4) reduces to

$$\prod_{i=1}^{n}[f(T_i)]^{\delta_i}[\bar{F}(S_i)]^{1-\delta_i}, \tag{5}$$

which is the likelihood for randomly right-censored data. These reductions follow from the fact that the model (3) leading to the likelihood (4) is more general than the usual censoring models. It may be noted that Mirzaei et al. (2015) has recently used the likelihood (4) for the purpose of parametric estimation of F.

3 Identifiability of time-to-event distribution

Before embarking on developing a method of estimation, we need to visit the issue of identifiability of the function of interest. Mirzaei et al. (2015) showed the likelihood (4) can also be interpreted as a product of conditional densities of (V_i, δ_i) given S_i, for $i = 1, 2, \ldots, n$, where $V_i = (S_i - T_i)\varepsilon_i\delta_i$. The density of $Y = (S, V, \delta)$ can be written as follows.

$$h(s, v, \delta) = \begin{cases} g(s)\bar{F}(s) & \text{if } v = 0 \text{ and } \delta = 0, \\ g(s)\int_0^s f(u)\pi(s-u)du & \text{if } v = 0 \text{ and } \delta = 1, \\ g(s)f(s-v)(1-\pi(v)) & \text{if } v > 0 \text{ and } \delta = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Here, g is the density of G (distribution of observation times), f is the density of the time-to-event (corresponding to the distribution function F) and π is the non-recall probability expressed as a function of the time elapsed since the occurrence of the event. Note that G can be a continuous, discrete or mixed distribution, and g represents its Radon-Nikodym derivative with respect to an appropriate dominating measure. The parameter of interest is the function F. We address the question as to whether the functions F, π and G are identifiable from h, in the next theorem.

Theorem 1 (a) *The distribution G is completely identifiable from h.*
(b) *If G has an absolutely continuous component over the support of F, then π and F are identifiable from h.*
(c) *If G has probability mass only over the space of integers and the function π comes from a family $\mathcal{P}$ satisfying the condition: '$\pi_1, \pi_2 \in \mathcal{P}$ implies that $(1-\pi_2)/(1-\pi_1)$ is not periodic with period one', then π and F are identifiable from h.*

The following example shows that if the condition given in part (c) of the above theorem does not hold, then f may not be identifiable from h.

Example 1 Let π_1 be a periodic function with period one, defined over the interval (0,1] by the equation $\pi_1(v) = v$, and $\pi_2 = 0.5$. Let $f_1(v) = 1/(t_{max} - t_{min})$ for $v \in [t_{min}, t_{max}]$, and $f_2 = f_1\,((1-\pi_1)/(1-\pi_2))$, defined over the same interval. Let g be any probability mass function defined over the space of positive integers. It may be verified that either of the triplets of functions (g, f_1, π_1) and (g, f_2, π_2), when substituted in (6), produce the following h:

$$h(1, v, \delta) = \begin{cases} g(s)(1 - \frac{s}{t_{max}-t_{min}}) & \text{if } v = 0 \text{ and } \delta = 0, \\ g(s)\frac{0.5s}{t_{max}-t_{min}} & \text{if } v = 0 \text{ and } \delta = 1, \\ g(s)\frac{1-(v-[v])}{t_{max}-t_{min}} & \text{if } v > 0 \text{ and } \delta = 1, \\ 0 & \text{otherwise.} \end{cases}$$

We now proceed with the problem of estimation, after assuming that either of the conditions given in part (b) and (c) of Theorem 1 are satisfied.

4 Nonparametric estimation

4.1 Reduction of the problem

It is known that nonparametric maximization of the likelihood (5) leads to the Kaplan-Meier estimator (Kaplan and Meier 1958), while maximization of (2) or (1) produces the Turnbull estimator (Turnbull 1976) or a special case of it. On the other hand, the likelihood (4) is difficult to maximize because of the integral contained in the expression. In order to simplify it, we assume that π is a piecewise constant function of the form

$$\pi(x) = b_1 I(x_1 < x \le x_2) + b_2 I(x_2 < x \le x_3) + \ldots + b_k I(x_k < x < \infty), \quad (7)$$

where $0 = x_1 < x_2 < \cdots < x_k$; $0 < b_1, b_2, \ldots, b_k \le 1$. Note that it is possible to constrain the parameters $b_1, b_2, \ldots, b_k$ to be in increasing order, so that π is a non-decreasing function. Such a choice correspond to the general perception that memory fades with time. However, we do not use this constraint in this paper.

When (7), the likelihood (4) reduces to

$$L = \prod_{i=1}^{n} \Bigg[\left\{ \sum_{l=1}^{k} b_l \big(F(W_l(S_i)) - F(W_{l+1}(S_i)) \big) \right\}^{1-\varepsilon_i} \left\{ f(T_i) \left(1 - \sum_{l=1}^{k} b_l I \big(W_{l+1}(S_i) < T_i \le W_l(S_i) \big) \right) \right\}^{\varepsilon_i} \Bigg]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \tag{8}$$

where $W_l(S_i) = (S_i - x_l) \vee t_{min}$ for $l = 1, \ldots, k$ and $W_{k+1}(S_i) = t_{min}, i = 1, 2, \ldots, n$. Note that

$$W_{k+1}(S_i) \le W_k(S_i) \le W_{k-1}(S_i) \le \cdots \le W_1(S_i) \le t_{max}, \tag{9}$$

Depending on the value of S_i, some of the above inequalities may in fact be equalities. Specifically, if l is an index such that $S_i - x_{l+1} \le t_{min} < S_i - x_l$, then $t_{min} = W_{k+1}(S_i) = \cdots = W_{l+1}(S_i)$. Further, if l is an index such that $S_i - x_{l+1} < t_{max} \le S_i - x_l$, then we have $W_l(S_i) = \cdots = W_1(S_i) = t_{max}$. The remaining equalities would be strict.

Anticipating point masses at T_i whenever $\delta_i \varepsilon_i = 1$, the likelihood (8) can be rewritten as

$$L = \prod_{i=1}^{n} \left[\left\{ \sum_{l=1}^{k} b_l \int_{A_{il}} f(u) du \right\}^{1-\varepsilon_i} \left\{ f(T_i) \left(1 - \sum_{l=1}^{k} b_l I \big(T_i \in A_{il} \big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}, \tag{10}$$

where $A_{il} = (W_{l+1}(S_i), W_l(S_i)]$ for $1 \le l < k$ and $A_{il} = [t_{min}, W_l(S_i)]$ for $l = k+1$. Note from (9) that some of the A_{il}'s can be empty. The simple form of the above likelihood, which can also be written without any integral, paves the way for estimation.

Assuming that k and $x_1, x_2, \ldots, x_k$ are known (while $b_1, b_2, \ldots, b_k$ are unknown), we attempt to narrow down the domain over which the likelihood needs to be maximized. Recall from Sect. 2 that for retrospective data with possible non-recall, the actual time of the landmark event of the ith subject can be said to belong to a set A_i. In case (i), A_i is an interval. In case (ii), it is a singleton set. In case (iii), it may be written as a union of the disjoint intervals used in the likelihood (10), i.e., $A_i = \bigcup_{l=1}^{k} A_{il}$. In summary, if we regard a singleton set as a special case of an interval, then each of the sets $A_i, i = 1, \ldots, n$, can be said to be constituted of a union of one or more intervals.

Consider the collection of all these intervals. Let $\{B_1, \ldots, B_M\}$ be the set of all unique members of this collection, disregarding replications. Following Turnbull (1976), we look for intersections of these intervals. Denote the non-empty subsets of the index set $\{1, 2, \ldots, M\}$ by $s_1, s_2, \ldots, s_{2^M-1}$. Define

$$I_r = \left\{ \bigcap_{i \in s_r} B_i \right\} \bigcap \left\{ \bigcap_{i \notin s_r} B_i^c \right\} \qquad \text{for } r = 1, 2, \ldots, 2^M - 1.$$

Some of the I_r's may be empty sets, denoted here by ϕ. Let

$$\mathcal{C} = \left\{ s_r : I_r \neq \phi, \ 1 \leq r \leq 2^M - 1 \right\}. \tag{11}$$

It can be verified that all the I_r's are distinct and disjoint. Let $\mathcal{A}$ be the set of intervals I_r such that $s_r \in \mathcal{C}$. Let $p_r = P(I_r)$, for all $I_r \in \mathcal{A}$. By using the definition of I_r, we can rewrite the contribution of individual i to the likelihood in the three cases of censoring as follows.

Case (i): Let $\delta_i = 0$. If l_i is the index such that $B_{l_i} = A_i$, then

$$P(A_i) = P(B_{l_i}) = \sum_{\substack{r : l_i \in s_r \\ s_r \in \mathcal{C}}} p_r.$$

Case (ii): Let $\delta_i \epsilon_i = 1$. If l_i is the index such that $I_{l_i} = A_i$ then $P(A_i) = p_{l_i}$.

Case (iii): Let $\delta_i (1 - \epsilon_i) = 1$. If $l_{i1}, l_{i2}, \ldots, l_{ik}$ are indices such that $B_{l_{it}} = A_{it}$ for $1 \leq t \leq k$, then

$$P(A_{it}) = P(B_{l_{it}}) = \sum_{\substack{r : l_{it} \in s_r \\ s_r \in \mathcal{C}}} p_r.$$

When the likelihood (10) is written in terms of the p_r's, it reduces to

$$L = \prod_{i=1}^{n} \left[\left[\left\{ \sum_{t=1}^{k} b_t \left(\sum_{\substack{r : l_{it} \in s_r \\ s_r \in \mathcal{C}}} p_r \right) \right\}^{1-\varepsilon_i} \left\{ p_r \Big|_{I_r = A_i} \left(1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} \left[\sum_{\substack{r : l_i \in s_r \\ s_r \in \mathcal{C}}} p_r \right]^{1-\delta_i} \right], \tag{12}$$

Thus, maximizing the likelihood (10) is equivalent to maximizing the likelihood (12) with respect to p_r for $s_r \in \mathcal{C}$.

There is a partial order among the members of $\mathcal{C}$ in the sense that some sets are contained in others. Let

$$\mathcal{C}_0 = \left\{ s_j : s_j \in \mathcal{C}; s_j \subset s_{j'} \text{ does not hold for any } s_{j'} \in \mathcal{C} \right\}. \tag{13}$$

Thus, $\mathcal{C}_0$ is a sub-class of $\mathcal{C}$ that retains only those sets that are *not* proper subsets of any other set. Our next result shows that the maximization of the likelihood can be restricted to this smaller class.

Theorem 2 *Maximizing the likelihood* (12) *with respect to p_r for $s_r \in \mathcal{C}$ is equivalent to maximizing it with respect to p_r for $s_r \in \mathcal{C}_0$, i.e.,*

$$\max_{p_r : p_r \in [0,1], \sum_{s_r \in \mathcal{C}} p_r = 1} L = \max_{p_r : p_r \in [0,1], \sum_{s_r \in \mathcal{C}_0} p_r = 1} L.$$

It follows from the above theorem that the likelihood has the same maximum value whether s_r is chosen from the class $\mathcal{C}$ or $\mathcal{C}_0$. Therefore, we can replace $\mathcal{C}$ by $\mathcal{C}_0$ in (12).

Let $\mathcal{A}_0$ be the set of distinct intervals I_j such that $s_j \in \mathcal{C}_0$. In order to simplify the notation, let $\mathcal{A}_0 = \{J_1, J_2, \ldots, J_m\}$ and $q_j = P(J_j)$. Let $\boldsymbol{p} = (q_1, q_2, \ldots, q_m)^T$ and $\boldsymbol{b} = (b_1, \ldots, b_k)^T$. Theorem 2 implies that the problem of maximizing (12) reduces to maximizing

$$\begin{aligned} L(\boldsymbol{p}, \boldsymbol{b}) &= L(q_1, \ldots, q_m, b_1, \ldots, b_k) \\ &= \prod_{i=1}^{n} \left[\left\{ \sum_{t=1}^{k} b_t \left(\sum_{j : J_j \subset A_{it}} q_j \right) \right\}^{1-\varepsilon_i} \left\{ \sum_{j : J_j = A_i} q_j \left(1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} \\ &\qquad \left[\sum_{j : J_j \subset A_i} q_j \right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \beta_{ij} q_j \right], \end{aligned} \tag{14}$$

subject to $\sum_{j=1}^{m} q_j = 1$, $0 \le q_1, \ldots, q_m \le 1$, and $0 \le b_1 \le \cdots b_k \le 1$, where

$$\beta_{ij} = \begin{cases} I\big(J_j \subseteq A_i\big) & \text{if } \delta_i = 0, \\ 1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) I\big(J_j \subseteq A_i\big) & \text{if } \delta_i \epsilon_i = 1, \\ \sum_{t=1}^{k} b_t I\big(J_j \subseteq A_{it}\big) & \text{if } \delta_i (1 - \epsilon_i) = 1, \end{cases} \tag{15}$$

for $i = 1, \ldots, n$, and $j = 1, \ldots, m$.

The task of identifying a maximum of the above likelihood is simplified through the following result, which is interesting by its own right.

Theorem 3 *Let $\mathcal{A}_1$ be the collection of singleton sets consisting of the exactly recalled times of events. Then $\mathcal{A}_1 \subseteq \mathcal{A}_0$. Further, if G is a discrete distribution with finite support, then the probability of $\mathcal{A}_0$ being equal to $\mathcal{A}_1$ goes to one, as $n \to \infty$.*

Let $\mathcal{I}_2$ be set of indices i (between 1 and n) that satisfy the conditions $\delta_i \varepsilon_i = 1$ with cardinality n_2. We are now ready for the next result regarding the existence and uniqueness of the NPMLE.

Theorem 4 *The likelihood* (14) *has a maximum. Further, if G is a discrete distribution with finite support, then the probability that* (14) *has a unique maximum goes to one, as* $n_2 \to \infty$.

4.2 A self-consistency algorithm

Following the work of Efron (1967) on computing the Kaplan-Meier estimator (Kaplan and Meier 1958) through a self consistency algorithm and similar work by Turnbull (1976) in the case of interval censored data, we seek to obtain an estimator of $\boldsymbol{p}$ for fixed $\boldsymbol{b}$, based on the self consistency approach.

For $i = 1, 2, \ldots, n$, let

$$L_{ij} = \begin{cases} 1 \text{ if } T_i \in J_j, \\ 0 \text{ otherwise,} \end{cases}$$

When $\delta_i \epsilon_i = 1$, the value of L_{ij} is known. Otherwise, its expectation with respect to the probability vector $\boldsymbol{p}$ is given by

$$E(L_{ij}) = \frac{\beta_{ij} q_j}{\sum_{j=1}^{m} \beta_{ij} q_j} = \mu_{ij}(\boldsymbol{p}), \quad \text{say.} \tag{16}$$

Thus, $\mu_{ij}(\boldsymbol{p})$ represents the probability that the i-th observation lies in J_j. The average of these probabilities across the n individuals,

$$\frac{1}{n} \sum_{i=1}^{n} \mu_{ij}(\boldsymbol{p}) = \pi_j(\boldsymbol{p}), \quad \text{say,} \tag{17}$$

should indicate the probability of the interval J_j. Thus, it is reasonable to expect that the vector $\boldsymbol{p}$ would satisfy the equation

$$q_j = \pi_j(\boldsymbol{p}) \quad \text{for} \quad 1 \le j \le m. \tag{18}$$

An estimator of $\boldsymbol{p}$ may be called self consistent if it satisfies the simultaneous equations (18).

The form of the above equations suggests the following iterative procedure.

Step I. Obtain a set of initial estimates $q_j^0 (1 \le j \le m)$.

Step II. At the nth stage of iteration, use current estimate, $\boldsymbol{p}^n$, to evaluate $\mu_{ij}(\boldsymbol{p}^n)$ for $i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, m$ and $\pi_j(\boldsymbol{p}^n)$ for $j = 1, 2, \ldots, m$ from (16) and (17), respectively.

Step III. Obtain improved estimates $\boldsymbol{p}^{n+1}$ by setting $q_j^{n+1} = \pi_j(\boldsymbol{p}^n)$.

Step IV. Return to Step II with $\boldsymbol{p}^{n+1}$ replacing $\boldsymbol{p}^n$.

Step V. Iterate; stop when the required accuracy has been achieved.

The following theorem shows that equation (18) defining a self consistent estimator must be satisfied by an NPML estimator of $\boldsymbol{p}$.

Theorem 5 *An NPML estimator of* $\boldsymbol{p}$ *must be self consistent.*

The above theorem may be proved by using a standard argument used in (Sun 2006, Sect. 3.4.1) for proving a similar result. For details, the reader is referred to Mirzaei and Sengupta (2013).

Let $\hat{\boldsymbol{p}} = (\hat{q}_1, \ldots, \hat{q}_m)$ and $\hat{\boldsymbol{b}} = (\hat{b}_1, \ldots, \hat{b}_k)$ denote values of $\boldsymbol{p}$ and $\boldsymbol{b}$, respectively, for which $L(\boldsymbol{p}, \boldsymbol{b})$ attains its maximum over the set

$$\Re = \left\{ (\boldsymbol{p}, \boldsymbol{b}) | \sum_{j=1}^{m} q_j = 1, \quad 0 \le q_1, \ldots, q_m \le 1, \quad 0 \le b_1 \le \cdots \le b_k \le 1 \right\}.$$

Then a maximum likelihood estimate $\hat{F}_n$ of F is given by

$$\hat{F}_n(t) = \sum_{j: J_j \subseteq [0,t]} \hat{q}_j. \tag{19}$$

In the sequel, we refer to this estimator as an NPMLE of F.

4.3 A computationally simpler estimator

The computational complexity of the NPMLE depends on the number of segments (k) used in the piecewise constant formulation of the function π. It follows from equations (11) and (13) that the cardinality of the class $\mathcal{C}$ can increase exponentially with k, though the cardinality of the sub-class $\mathcal{C}_0$ is smaller. One can conceive of a computational simplification on the basis of Theorem 3. According to this theorem, the NPMLE has mass only at points of exact recall of the event, when n is large. In such a case, the likelihood (14) involves J_j's that are singletons only. Hence, the crucial task of identifying the appropriate J_j's becomes redundant. Therefore, irrespective of the value of n, one can maximize (14) with respect to point masses restricted to the time points of exact recall of the event. This method would produce a computationally simpler estimate that is equivalent to the unique NPMLE for large n.

Formally, let $t_1, \ldots, t_p$ be the ordered set of distinct ages at event that have been perfectly recalled, and $q_1^*, \ldots, q_p^*$ be the probability masses allocated to them. The likelihood (14), subject to the constraint that $q_j = 0$ whenever $J_j \notin \mathcal{A}_1$, is equivalent to the unconstrained maximization of

$$L(q_1^*, \ldots, q_p^*, \boldsymbol{b}) = \prod_{i=1}^{n} \left[\sum_{j=0}^{p} \alpha_{ij} q_j^* \right], \tag{20}$$

where

$$\alpha_{ij} = \begin{cases} 1 & \text{if } t_j \in A_i,\ \delta_i = 0, \\ 1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) & \text{if } t_j \in A_i,\ \delta_i \epsilon_i = 1, \\ \sum_{t=1}^{k} b_t I(t_j \in A_{it}) & \text{if } \delta_i(1-\epsilon_i) = 1, \end{cases} \tag{21}$$

with respect to the parameters $(q_1^*, \ldots, q_p^*)$ and $\boldsymbol{b}$, over the set

$$\Re^* = \left\{ (q_1^*, \ldots, q_p^*, \boldsymbol{b}) | \sum_{j=1}^{p} q_j^* = 1, \quad 0 \le q_1^*, \ldots, q_p^* \le 1,\ 0 \le b_1 \le \cdots \le b_k \le 1 \right\}.$$

Let the likelihood (20) be maximized at $(\hat{q}_1^*, \ldots, \hat{q}_p^*, \hat{\boldsymbol{b}}^*)$. We define an approximate NPMLE (AMLE) of F as

$$\tilde{F}_n(t) = \sum_{j:t_j \le t} \hat{q}_j^*. \tag{22}$$

4.4 Estimation of variance

The variance of the NPMLE and the AMLE may be estimated through bootstrap resampling. Sen et al. (2010) have argued that the usual bootstrap is not be guaranteed to be consistent. The variances of (19) and (22) may be estimated through m out of n bootstrapping of Bickel et al. (1997) with selection of m as in Bickel and Sakov (2008), so that consistency is ensured.

5 Consistency of the estimators

Consider the set-up of Sect. 2 and the estimator $\tilde{F}_n$. Let Θ be the set of all distribution functions over the support $[t_{min}, t_{max}]$, i.e.,

$$\Theta = \Big\{ F\ : [t_{min}, t_{max}] \to [0,1];\ F \text{ rightcontinuous, nondecreasing; } F(t_{min}) = 0;\ F(t_{max}) = 1 \Big\},$$

and $\overline{\Theta}$ be the set of all sub-distribution functions, i.e.,

$$\overline{\Theta} = \Big\{ F\ : [t_{min}, t_{max}] \to [0,1];\ F \text{ rightcontinuous, nondecreasing; } F(t_{min}) = 0;\ F(t_{max}) \le 1 \Big\}.$$

Note that, with respect to the topology of vague convergence, $\overline{\Theta}$ is compact by Helley's selection theorem. Further, let F_0 denote the true distribution of the time of occurrence of landmark events, and $F_0(t_{min}) = 0$. Let the interview times assume values over set $\mathcal{S}$.

For any given distribution $F \in \Theta$ having masses restricted to the set $\{t_1, \ldots, t_p\}$, the log of the likelihood (20) can be written as

$$\ell(F) = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{p} \alpha_{ij} \left\{ F(t_j) - F(t_{j-}) \right\} \right]. \tag{23}$$

Define the set

$$\mathcal{E} = \{F \, : \, F \in \Theta, \; E[\ell(F) - \ell(F_0)] = 0\}, \tag{24}$$

which is an equivalence class of the true distribution F_0.

Strong consistency of the AMLE is established by the following theorem.

Theorem 6 *In the above set-up, the AMLE $\{\tilde{F}_n\}$ converges almost surely to the equivalence class $\mathcal{E}$ of the true distribution F_0, in the topology of vague convergence.*

The following theorem establishes consistency of the NPMLE.

Theorem 7 *In the set-up described before Theorem 6, the NPMLE $\{\hat{F}_n\}$ converges in probability to the equivalence class $\mathcal{E}$ of the true distribution F_0, in terms of the Lévy distance.*

The last theorem of this section ensures that under some conditions the equivalence class used in Theorems 6 and 7 includes only F_0.

Theorem 8 *If either the condition given in part (b) of Theorem 1 or the pair of conditions given in part (c) holds, then the equivalence class defined in (24) is the singleton class $\{F_0\}$.*

6 Simulations

For the purpose of simulation, we generate sample times to landmark event from the Weibull distribution with shape and scale parameters $\alpha = 11$ and $\beta = 13$, respectively, and truncate the generated samples to the interval [8,16]. This truncated distribution has median of 11.57. The corresponding 'time of interview' is generated from the discrete uniform distribution over $\{7, 8, \ldots, 21\}$. These choices are in line with the data analytic example of the next section, where the time to landmark event is the age at menarche in years. As for the non-recall probability, we use (7) with $k = 8$, intervals of equal length and three sets of values of the parameters described in Table 1.

Case (a) corresponds to rapid and extensive forgetting with the passage of time, while Case (b) represents better retention. The choice of constant π function in Case (c) makes the censoring non-informative. Case (a) should favour the proposed methods, as the chosen function π induces informative censoring. Case (c) is ideal for the Turnbull estimator based on censored duration data, as the censoring is non-informative, while the proposed estimators are burdened with unnecessary nuisance parameters. Case (b) may not favour any method decisively, as the non-recall probability, though informative, is relatively small and consequently the informativeness of the censoring is mild. The proposed methods, on the other hand, have the handicap of nuisance parameters.

The NPMLE and AMLE of F are implemented by assuming that $k, x_1, x_2, \ldots, x_k$ in (7) are known, while $b_1, b_2, \ldots, b_k$ are estimated. The NPMLE and the AMLE

Table 1 Values of $b_1, \ldots, b_8$ in three simulation models and resulting proportion of data with different types of incompleteness

Simulation model	Case (a)	Case (b)	Case (c)
Value of b_1 and b_2	0.1	0.05	0.40
Value of b_3	0.40	0.15	0.40
Value of $b_4, \ldots, b_8$	0.95	0.35	0.40
Percentage of cases with $\delta_i = 0$	39%	39%	39%
Percentage of cases with $\delta_i \varepsilon_i = 1$	36%	51%	27%
Percentage of cases with $\delta_i (1 - \varepsilon_i) = 1$	25%	10%	34%

are obtained by maximizing the likelihoods (14) and (20), respectively. Recursive maximization is carried out alternately with respect to the probability parameters and $\boldsymbol{b}$. Since k is chosen as 8, there are eight different b_t's (nuisance parameters) to be estimated along with NPMLE and AMLE, even though many of the b_t's have equal values.

We compare the performances of the NPMLE (19) and the AMLE (22) with the two MLEs based on (1) and (2), described here as the Turnbull estimator (status) and the Turnbull estimator (duration), respectively. As a benchmark, we also evaluate the performance of the empirical distribution function (EDF), a hypothetical estimator computed from the underlying complete data. The results reported here are based on 500 simulation runs for sample sizes $n = 100$, 300 and 1000. The simulations for the three cases are run parallelly. For each run, the complete data as well as the observation times for the three cases are the same, while the events of forgetting are simulated subsequently according to the chosen non-recall probability.

The Turnbull estimator (status) is uniquely defined only at integer ages. Therefore, in all the plots, we represent it through a set of unconnected points at integer ages.

Figure 1 shows plots of the bias, the variance and the mean square error (MSE) of the five estimators for different ages, for $n = 100$ and parameters of the non-recall probability function (7) chosen as in Case (a). The NPMLE is found to have smaller bias than the Turnbull estimator (duration), and smaller variance than the Turnbull estimator (status), and smaller MSE than both the Turnbull estimators. The bias of the AMLE is only marginally worse than that of NPMLE, and the MSE is comparable.

Figure 2 shows these plots for $n = 100$ and parameters of the non-recall probability function (7) chosen as in Case (b). Even though the bias of the estimators reduce, the overall pattern of performances remains the same. The Turnbull estimator (duration) appears to have smaller bias when forgetting is less prevalent. The performance of the AMLE is almost identical to that of NPMLE. The similarity of performances of the Turnbull estimator (duration), the NPMLE and the AMLE may be explained by the fact that the nature of treatment of the cases with forgotten dates of events matters less when there is less forgetting. The EDF and the Turnbull estimator (status) have exactly the same performance as depicted in Fig. 1, since the data required for these estimators remain unchanged.

Figure 3 shows these plots for $n = 100$ and parameters of the non-recall probability function (7) chosen as in Case (c). The performances of the EDF and the Turnbull

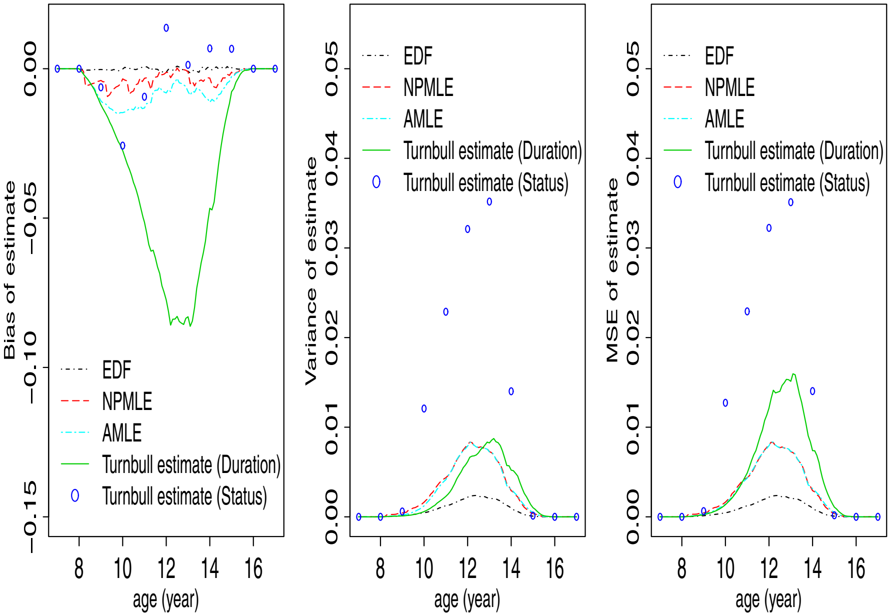

Fig. 1 Comparison of bias, variance and MSE of the four estimator in case (a) and $n = 100$

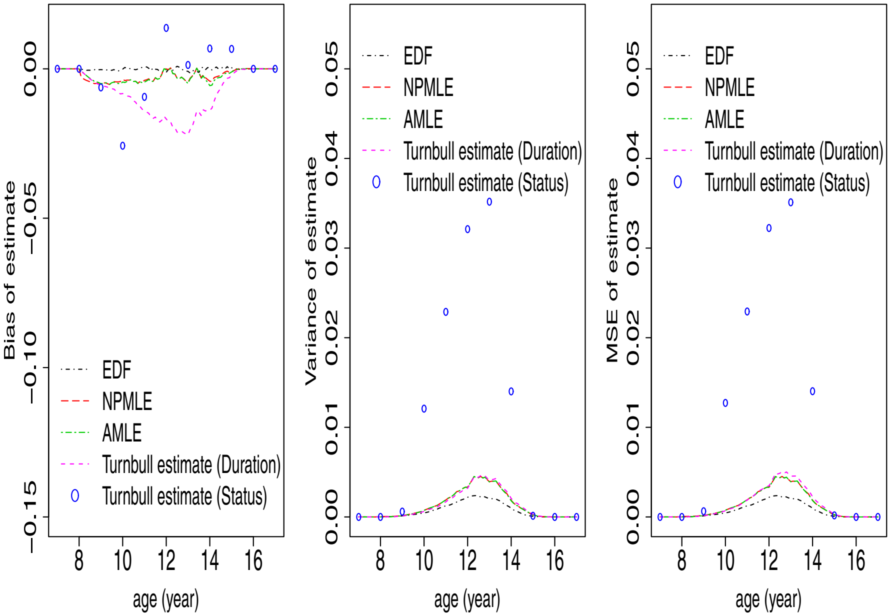

Fig. 2 Comparison of bias, variance and MSE of the four estimator in case (b) and $n = 100$

estimator (status) continue to be as seen in Figs. 1 and 2. The Turnbull estimator (duration), the NPMLE and the AMLE have similar patterns of bias, variance and MSE. This is noteworthy, as the constancy of the non-recall probability works in

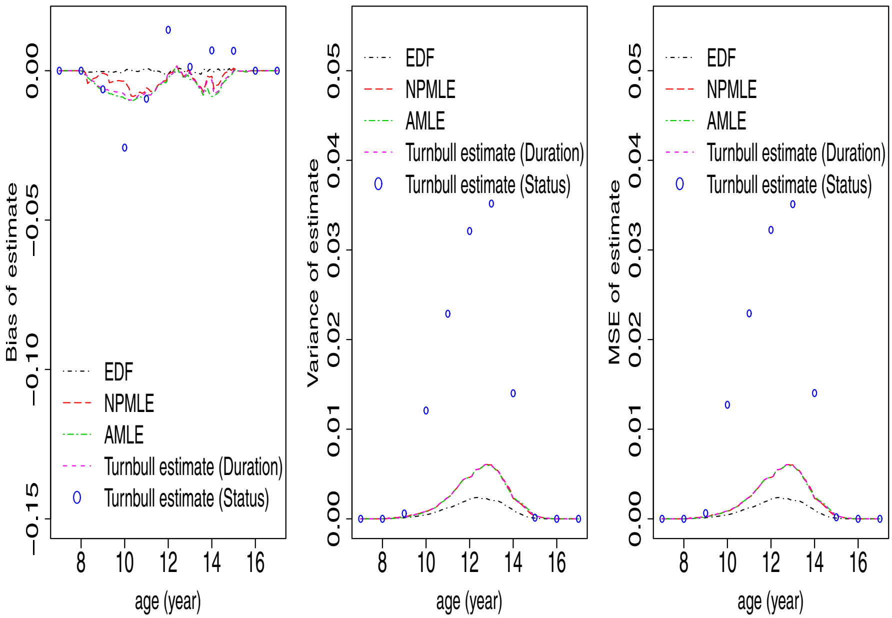

Fig. 3 Comparison of bias, variance and MSE of the four estimator in case (c) and $n = 100$

favour of the Turnbull estimator (duration), which does not involve estimation of the nuisance parameters $b_1, \ldots, b_8$.

In all the cases, the performances of the Turnbull estimator (duration), the NPMLE and the AMLE are noticeably worse than that of the EDF. This is because of the substantial number of right censored observations (with $\delta_i = 0$), as seen from Table 1. The superior performance of the NPMLE and the AMLE in comparison with the Turnbull (status) shows how usefully recall data can be utilized.

Figures 4, 5 and 6 show plots similar to Figs. 1, 2 and 3 for $n = 300$. There is a marked reduction in the bias and variance of the NPMLE, the AMLE and the Turnbull estimator (status). The previously observed pattern of relative performances continues to prevail. The bias of the Turnbull estimator (duration) observed in Fig. 6 is smaller in comparison with the same case with $n = 100$ (Fig. 3). This is expected, as the interval censoring associated with forgetting the date of event is chosen to be non-informative in this case. There is no such reduction in Figs. 4 and 5 though. The patterns of bias of the Turnbull (duration) estimator observed in these two figures are of the same order as observed in Figs. 1 and 2 respectively. This occurrence underscores the cost of inadequate handling of the cases of non-recall.

Simulations for $n = 1,000$ in Cases (a), (b) and (c), leading to Figs. 7, 8 and 9, show that the bias and the variance of the Turnbull estimator (status), the NPMLE and the AMLE continue to reduce with sample size. The same can be said about the Turnbull estimator (duration) in Case (c), as observed from Fig. 9. In contrast, the bias of the Turnbull estimator (duration) appears to have settled at a value away from 0, in Cases (a) and (b), as observed in Figs. 7 and 8.

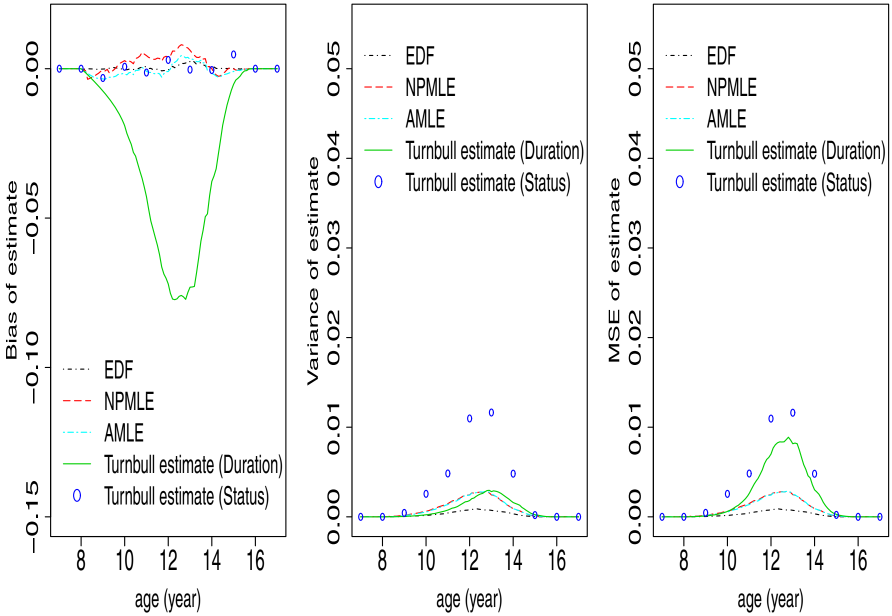

Fig. 4 Comparison of bias, variance and MSE of the four estimator in case (a) and $n = 300$

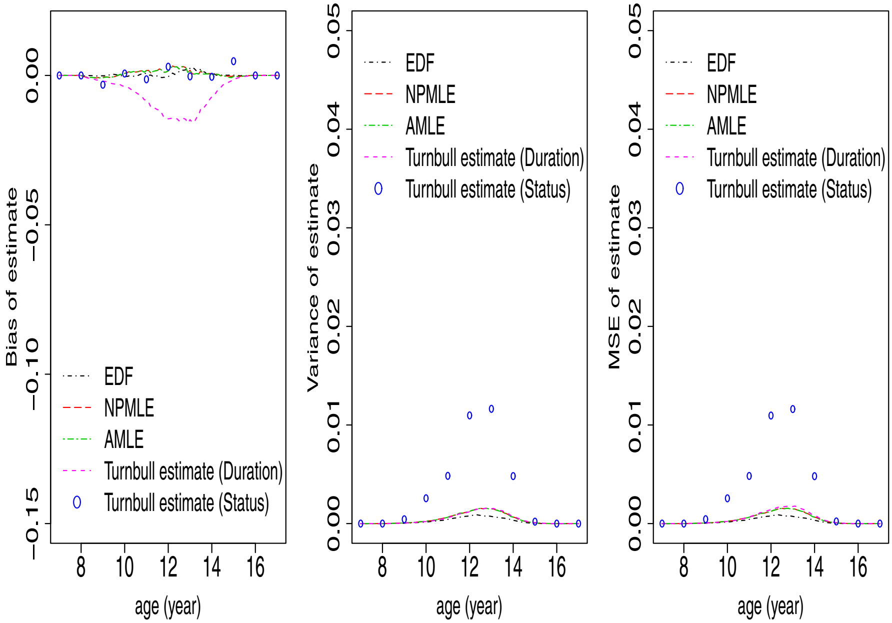

Fig. 5 Comparison of bias, variance and MSE of the four estimator in case (b) and $n = 300$

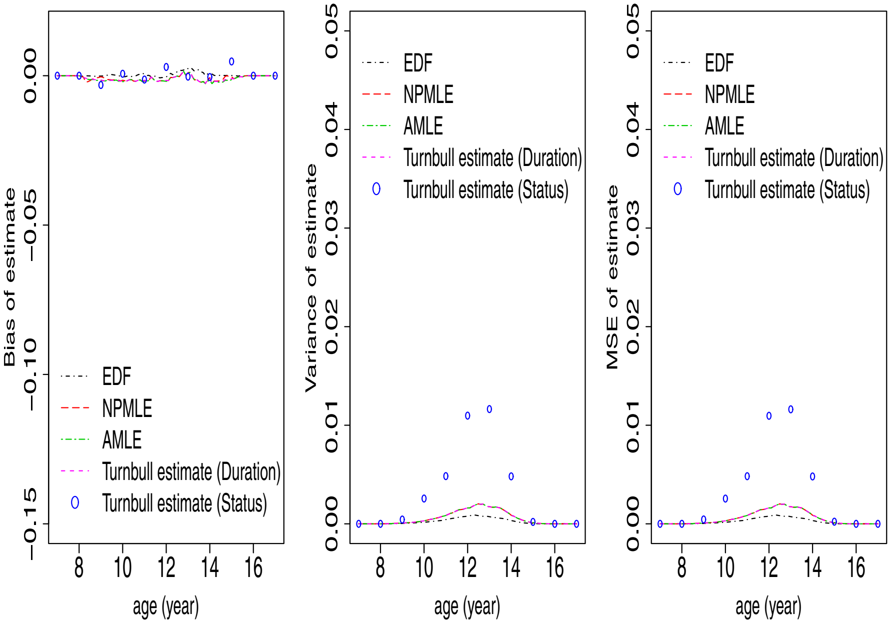

Fig. 6 Comparison of bias, variance and MSE of the four estimator in case (c) and $n = 300$

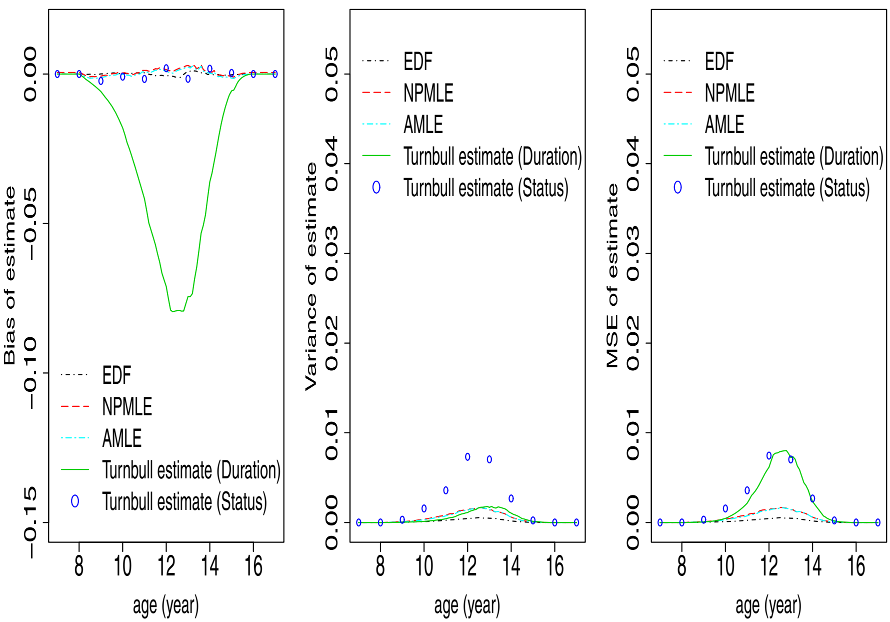

Fig. 7 Comparison of bias, variance and MSE of the four estimator in case (a) and $n = 1000$

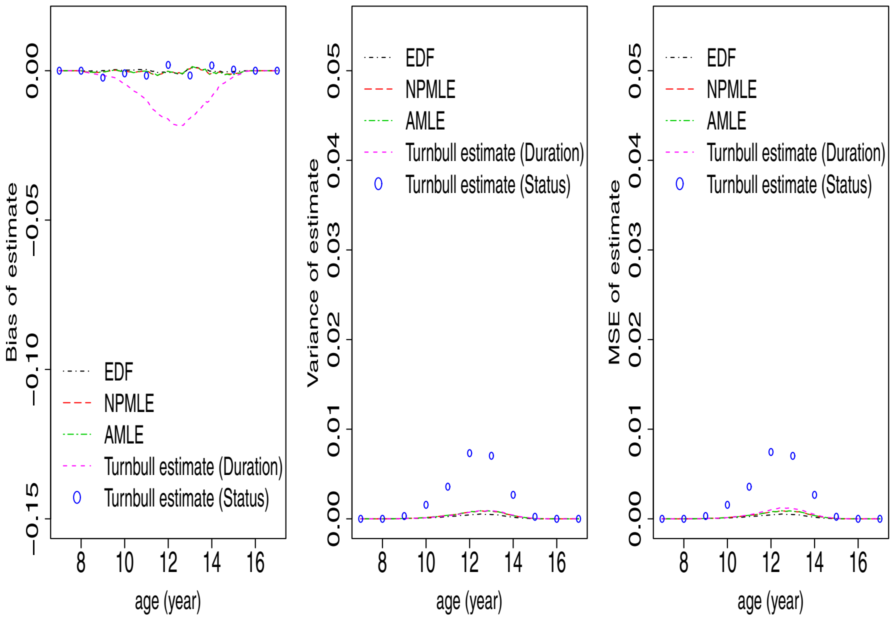

Fig. 8 Comparison of bias, variance and MSE of the four estimator in case (b) and $n = 1000$

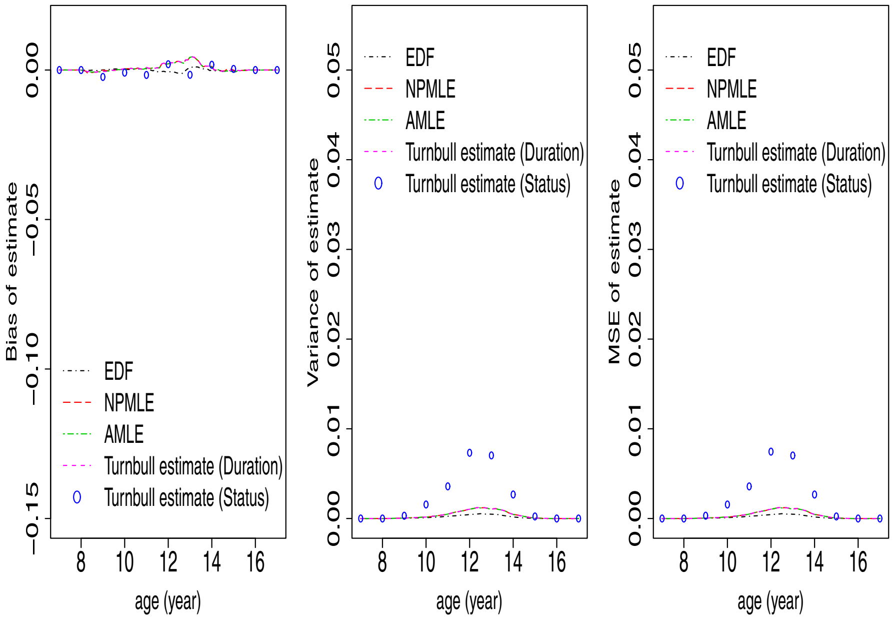

Fig. 9 Comparison of bias, variance and MSE of the four estimator in case (c) and $n = 1000$

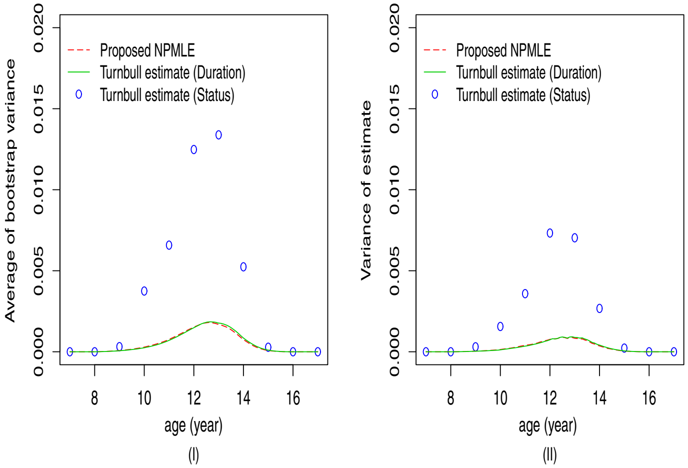

Fig. 10 **I** Average of bootstrap variance estimator and **II** Sample variance of the four estimators of F

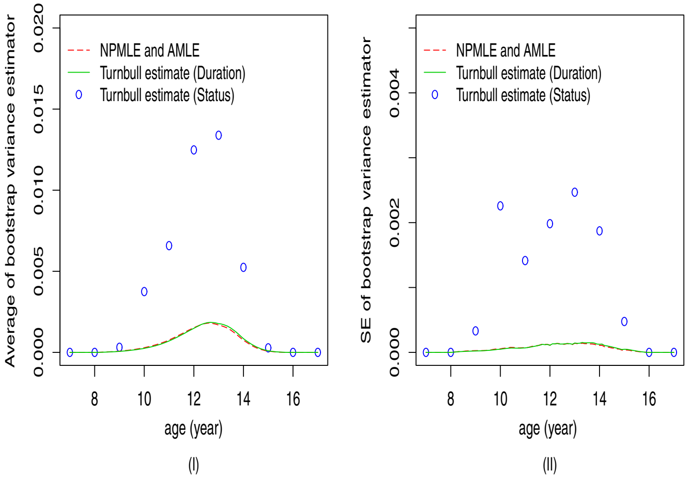

Fig. 11 **I** Average and **II** Standard error of bootstrap variance estimator using four methods

On the basis of the above findings, the AMLE may be regarded as a reasonable substitute for the NPMLE.

We now turn to the performance of the bootstrap estimator of variance. For this study, we choose $n = 1,000$ and the parameters of the non-recall probability function as in Case (b). We choose the m out of n bootstrap of Bickel et al. (1997), with $m = n^{0.8}$ (see Bickel and Sakov 2008). Figure 10 shows the plots of the average (across 500 runs) of the bootstrap estimate of variance of the NPMLE and the AMLE shown in panel (I) and the sample variance (across 500 runs) of the two estimators in panel (II). The corresponding plots for the other estimators are also shown. The two sets of the plots show comparable patterns, and mild overestimation of variance on the average. Figure 11 shows the standard error (across 500 runs) of the bootstrap estimator of variance, alongside the average (across 500 runs) of the same. It is seen that the standard error is generally much smaller than the average. Thus, the bootstrap estimator of variance appears to be a reasonable one.

Plots for other sample sizes and other values of parameters, which show similar patterns, are omitted for the sake of brevity.

7 Data analysis

In a recent anthropometric study conducted by the Biological Anthropology Unit of the Indian Statistical Institute in and around the city of Kolkata, India from 2005 to 2011 (ISI 2012, p. 108), a total of 2195 randomly selected individuals, aged between 7 and 21 years, were surveyed. The subjects were interviewed on or around their birthdays. The data set contains age, menarcheal status, age at menarche (if recalled), and some other information.

For this data set, the landmark event is the onset of menarche. Among 2195 samples, 775 individuals did not have menarche, 443 individuals had menarche and recall the date of its onset and 977 individuals had menarche but could not recall the date. We modeled the non-recall probability π, over the interval 0 to 13 years (maximum possible separation between menarcheal age and age at observation in the sample). We used a piecewise constant model, with $k = 8$ and a uniform grid. Figure 12 shows the NPMLE, the AMLE, the Turnbull estimator (duration) and the Turbull estimator (Status) of the distribution function of the age at menarche. It can be seen that the NPMLE and the AMLE are indistinguishable. The NPMLE, the AMLE and the Turnbull (status) estimator are closer to one another as compared to the Turnbull (duration) estimator, which is expected to be biased. Since the Turnbull estimator (status) is not uniquely defined at non-integer ages, the NPMLE or the AMLE may be preferred.

In order to get an idea about the estimation error, we estimate the variances of the NPMLE, the AMLE and the two Turnbull estimators through bootstrap resampling. As in the previous section, we use m out of n bootstrap of Bickel et al. (1997), with

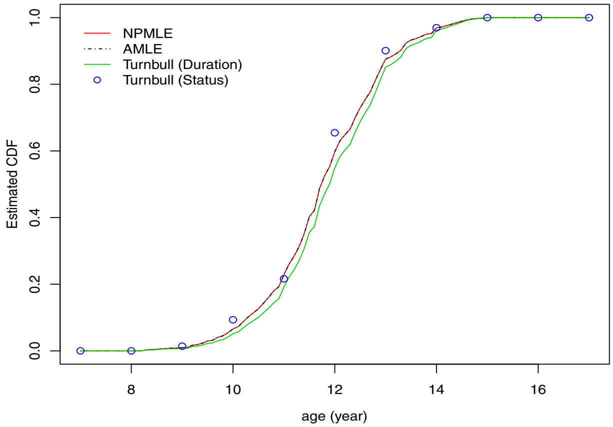

Fig. 12 Estimated distribution function of data using three methods

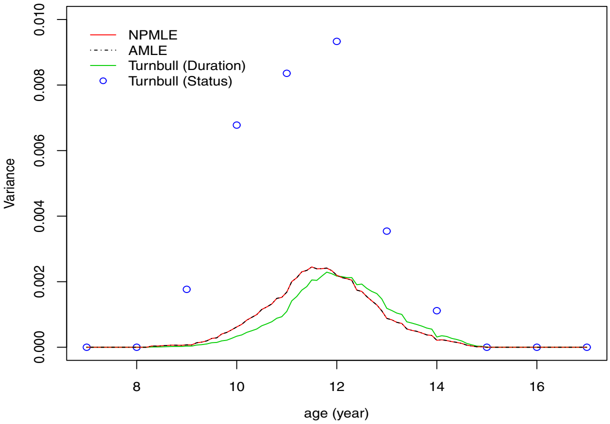

Fig. 13 Variance of estimated distribution function of data using three methods

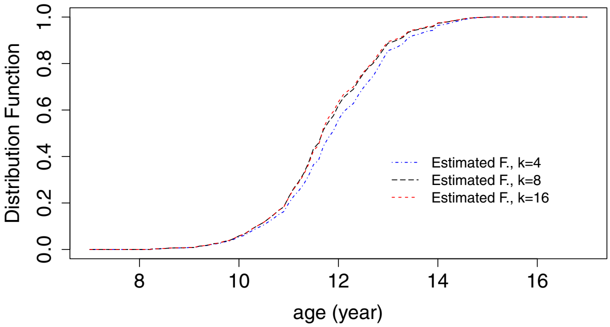

Fig. 14 Estimated distribution function with different k

$m = n^{0.8}$, i.e., $m = 472$, and 500 replications. Plots of the bootstrap estimators of variance of the three estimators, shown in Figure 13, reveal that the Turnbull estimator (status) has a much larger variance compared to the NPMLE and the AMLE. Thus, the NPMLE and the AMLE may be preferred. The computational simplicity of the AMLE makes it more attractive than the NPMLE.

The chosen value of k for estimation was obtained after considering a coarser and a finer partition for the piecewise constant model of π. Specifically, the range 0 to 13 years was split experimentally into k equal intervals, with $k = 4$, 8 and 16, and the resulting estimated distribution functions were compared. Figure 14 shows plots of the estimated distribution function for different values of k. It is seen that by increasing k from 4 to 8, one observes a substantial change in the estimated distribution function, though the change is much less when k is increased from 8 to 16. The integrated mean square difference between the distribution functions (scaled by the integral of the square of the function for the lower value of k) is 0.85 when one compares $k = 4$ with $k = 8$. The same criterion produces the value 0.019 when the comparison is between the curves for $k = 8$ and $k = 16$. We have chosen $k = 8$, as the alternative choice $k = 16$ does not produce a substantially different estimate of the distribution

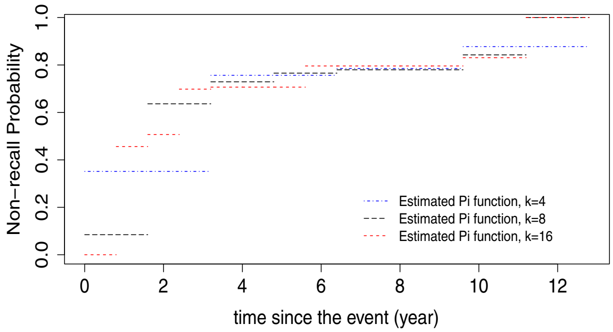

Fig. 15 Estimated π function with different k

function. Figure 15 shows the estimated function π for different values of k. Once again, the estimates of π for $k = 8$ and $k = 16$ differ much less than those for $k = 4$ and $k = 8$. This observation justifies the choice $k = 8$.

8 Concluding remarks

In this paper, we have offered a realistic model and method for estimating the time-to-event distribution based on recall data, in the presence of informative censoring. Modeling of the incompleteness in the data is a critical issue. There can be an alternative approach for modeling this kind of incomplete data, through an underlying distribution (F) for the time till the occurrence of the event of interest, and another distribution (say, G) for the time from that occurrence to the forgetting of the date. The latter may in fact be a sub-distribution function, with some mass at infinity. In this formulation also, there would be three cases for individual i: neither event has occurred till the age at interview ($\delta_i = 0$), only the first event has occurred ($\delta_i \epsilon_i = 1$) and both events have occurred ($\delta_i(1 - \epsilon_i) = 1$). The contribution of individual i to the likelihood in the three cases are as follows.

Case (i): When $\delta_i = 0$, the contribution of the individual in likelihood is $\bar{F}(S_i)$.

Case (ii): When $\delta_i \epsilon_i = 1$, the contribution of the individual in likelihood is $f(T_i)\bar{G}(S_i - T_i)$.

Case (iii): When $\delta_i(1 - \epsilon_i) = 1$, the contribution of the individual in likelihood is $\int_0^{S_i} f(u)G(S_i - u)du$.

It can be seen that these contributions also lead to the likelihood (4), with G replacing π. In fact, the above formulation provides an interpretation of the ‘forgetting function’ π as the distribution function of the time to the forgetting event, measured from the date of occurrence of the main event. This interpretation holds when π is non-decreasing, while the general formulation of Section 2 remains applicable even when π does not have this property.

The approach of modeling non-recall through a forgetting function may be adapted to the estimation of the distribution of the time from contracting HIV infection through blood transfusion to the onset of AIDS (Kalbfleisch and Lawless 1989). Here, the subjects listed in a central registry have a known date of onset of AIDS, but the date of transfusion is sometimes difficult to ascertain retrospectively. However, a range or set of dates may be available. If one ignores the issue of truncation, as in Kalbfleisch and Lawless (1989), then the non-recalled cases may perhaps be handled in a better way. Instead of ignoring these cases altogether, one may incorporate this censored information through modeling of recall uncertainty following the approach used in the paper. Incorporation of truncation would lead to a more complicated likelihood, and a different computational algorithm may have to be explored.

Acknowledgments This research is partially sponsored by the project 'Physical growth, body composition and nutritional status of the Bengal school aged children, adolescents, and young adults of Calcutta, India: Effects of socioeconomic factors on secular trends', funded by the Neys Van Hoogstraten Foundation of the Netherlands. The authors thank Professor Parasmani Dasgupta, leader of the project, for making the data available for this research. The interpretation given in the end of Section 8 emerged from a discussion with Professor John D. Kalbfleisch of the University of Michigan. The authors wish to thank him for helpful discussion. The authors thank two referees and an associate editor for making extensive and useful comments that led to substantial improvement of the manuscript.

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

Research involving Human Participants and/or Animals This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

Appendix

Proof of Theorem 1

(a) We have, from (6) (with $v > 0$ and $\delta = 1$),

$$h(s, v, 1) = g(s) f(s - v)(1 - \pi_\eta(v)),$$

that is,

$$1 - \pi_\eta(v) = \frac{h(s, v, 1)}{g(s) f(s - v)} \qquad \forall s, v \text{ s.t. } v < s. \tag{25}$$

By substituting the above expression in (6) for $v = 0$ and $\delta = 1$ and simplifying the equation, we have

$$F(s) = \frac{h(s, 0, 1) + \int_0^s h(s, s - u, 1) du}{g(s)}. \tag{26}$$

By substituting the above expression of $F(s)$ in (6) for $v = 0$ and $\delta = 0$, we have $g(s)$ as

$$g(s) = h(s, 0, 0) + h(s, 0, 1) + \int_0^s h(s, s - u, 1) du. \tag{27}$$

The above identity holds over the support of G irrespective of whether G is a discrete, a continuous or a mixed distribution. The identifiability of G follows.

(b) By substituting (27) in (26), we have

$$F(s) = \frac{h(s, 0, 1) + \int_0^s h(s, s-u, 1)du}{h(s, 0, 0) + h(s, 0, 1) + \int_0^s h(s, s-u, 1)du}. \tag{28}$$

If G has an absolutely continuous component over the support of F, then for every s and all real valued $v < s$, we have from (25),

$$\pi_\eta(v) = 1 - \frac{h(s, v, 1)}{g(s) f(s-v)}. \tag{29}$$

Thus, (29) together with (27) and (28) identify F and π_η completely.

(c) For the sake of contradiction, let us assume there are two pairs of choices of f and π_η, say (f_1, π_1) and (f_2, π_2), such that their substitution in the right hand side of (6) produces the same function. If we follow the steps leading to (25) for these two pairs of functions, then we have, for all integers s and all $v < s$,

$$f_1(s-v)(1-\pi_1(v)) = f_2(s-v)(1-\pi_2(v)).$$

Hence,

$$\frac{f_1(v)}{f_2(v)} = \frac{1-\pi_2(s-v)}{1-\pi_1(s-v)} \qquad \forall s, v \text{ s.t. } v < s. \tag{30}$$

Since the above identity holds for all integers s, we can write

$$\frac{1-\pi_2(s-v)}{1-\pi_1(s-v)} = \frac{1-\pi_2(1-v)}{1-\pi_1(1-v)} \quad \text{for all integer } s \text{ and all } v < s. \tag{31}$$

The above equation implies that the function $(1-\pi_1)/(1-\pi_2)$ is periodic over the relevant domain with period 1, which contradicts the assumption. Therefore, the pair (f, π_η) is uniquely defined for any given h.

Proof of Theorem 2

By definition of $\mathcal{C}$ and $\mathcal{C}_0$ we can rewrite the likelihood (12) as follows.

$$L = \prod_{i=1}^{n} \left[\left\{ \sum_{t=1}^{k} b_t \left(\sum_{\substack{r: l_{it} \in s_r \\ s_r \in \mathcal{C}_0}} p_r + \sum_{\substack{r: l_{it} \in s_r \\ s_r \in \mathcal{C} \setminus \mathcal{C}_0}} p_r \right) \right\}^{1-\varepsilon_i} \left\{ p_{l_i} \left(1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i}$$
$$\left[\sum_{\substack{r: l_i \in s_r \\ s_r \in \mathcal{C}_0}} p_r + \sum_{\substack{r: l_i \in s_r \\ s_r \in \mathcal{C} \setminus \mathcal{C}_0}} p_r \right]^{1-\delta_i}. \tag{32}$$

For any $s_r \in \mathcal{C}\backslash\mathcal{C}_0$, let $\mathcal{A}_r = \{I_{r'} : s_{r'} \in \mathcal{C}_0, s_r \subset s_{r'}\}$. By construction of $\mathcal{C}_0$, $\mathcal{A}_r$ is a non-empty set. The elements of $\mathcal{A}_r$ are disjoint sets consisting of unions of intervals, which are subsets of $[t_{min}, t_{max}]$. Let I_{r^*} be that member of $\mathcal{A}_r$ which satisfies the condition 'there is $\alpha \in I_{r^*}$ such that $\alpha < \beta$ whenever $\beta \in I_{r^\dagger}$ for any $I_{r^\dagger} \in \mathcal{A}_r$'. We shall show that by shifting mass from any I_r to I_{r^*}, there will be no reduction in the contribution of any individual to the likelihood (32).

We can check the effect of shifting mass on contribution of different individuals ($i = 1, \ldots, n$) to the likelihood.

Case (i). Let $\delta_i = 0$. If $l_i \in s_r$ or $l_i \notin s_{r^*}$, then there is no change in the likelihood. If $l_i \in s_{r^*}\backslash s_r$, then the factor contributed by individual i to the likelihood increases by p_r.

Case (ii). Let $\delta_i \varepsilon_i = 1$. If $l_i \notin s_{r^*}$, then there is no change in the likelihood. If $l_i \in s_{r^*}\backslash s_r$, then the factor contributed by individual i to the likelihood increases by p_r. The case $l_i \in s_r$ cannot occur, because I_r and I_{r^*} are distinct and disjoint.

Case (iii). Let $\delta_i(1 - \varepsilon_i) = 1$. There exists at most one t such that $l_{it} \in s_r$. If there is such a t, then there is no change in the likelihood. If there exists no t such that $l_{it} \in s_{r^*}$, then there is no change in the likelihood also. In case there is a t such that $l_{it} \in s_{r^*}\backslash s_r$, the factor contributed by individual i to the likelihood increases by p_r.

It follows that maximizing L can be restricted to $\{p_r : s_r \in \mathcal{C}_0\}$.

Proof of Theorem 3

It is easy to see, from the construction of $\mathcal{A}_0$, that every singleton set consisting of a perfectly recalled time-to-event is a nominal interval with zero width, belonging to $\mathcal{A}_0$. Therefore $\mathcal{A}_1 \subseteq \mathcal{A}_0$.

Define $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ as index sets of individuals in the three different cases of censoring. The interview times are discrete valued with finite domain; $x_1, x_2, \ldots, x_k$ are also finite. Therefore, even when n is large, there is at most a finite number (say N) of distinct sets of the form

$$A_s = \{\bigcap_{i \in s} B_i\} \bigcap \{\bigcap_{i \in \mathcal{S}_1 \cup \mathcal{S}_3 \backslash s} B_i^c\},$$

where $s \subseteq \mathcal{S}_1 \cup \mathcal{S}_3$. Denote $s^{(1)}, s^{(1)}, \ldots, s^{(N)}$, the index sets corresponding to the N distinct sets described above.

Consider a member of $\mathcal{A}_0$, say I_s, where s is a subset of $\{1, 2, \ldots, n\}$. If $s \subseteq \mathcal{S}_2$, then it is already a singleton. If not, it can be written as $s^{(j)} \cup (s \backslash s^{(j)})$, with $s^{(j)} \subseteq \mathcal{S}_1 \cup \mathcal{S}_3$ and $s \backslash s^{(j)} \subseteq \mathcal{S}_2$ for some $j \in \{1, 2, \ldots, N\}$. Let us consider three further special cases.

Case (a). Let $s = s^{(j)} \cup \{r\}$ for $r \in \mathcal{S}_2$. In this case, I_s is either a singleton or a null set. If it is a null set, then it cannot be a member of $\mathcal{A}$, and hence of $\mathcal{A}_0$. Thus, Case (a) contributes only singletons to $\mathcal{A}_0$.

Case (b). Let $s = s^{(j)} \cup \{r_1, r_2, \ldots, r_p\}$, for $r_1, r_2, \ldots, r_p \in \mathcal{S}_2$ when $p > 1$. In this case, I_s is either a singleton or a null set. Since the absolute continuity of the time-to-event distribution almost surely precludes coincidence of two sample values (say, T_{r_1} and T_{r_2}), I_s is a null set with probability 1. In summary, Case (b) cannot contribute anything other than a singleton to $\mathcal{A}_0$.

Case (c). Let $s = s^{(j)}$. The probability that a specific individual (say, the i-th one) has the landmark event at an age contained in $A_{s^{(j)}}$ is

$$P(T_i \in A_{s^{(j)}}, \delta_i \epsilon_i = 1).$$

Since this quantity is strictly positive, the probability that none of the n individuals have had the landmark event in $A_{s^{(j)}}$ and recalled the date is

$$\left(1 - P(T_i \in A_{s^{(j)}}, \delta_i \epsilon_i = 1)\right)^n,$$

which goes to zero as $n \to \infty$. Thus, the probability that there is $i \in \mathcal{S}_2$ such that $T_i \in A_{s^{(j)}}$ goes to one as $n \to \infty$. Therefore, $I_{s^{(j)} \cup \{i\}} = I_{s^{(j)}} \cap \{T_i\}$ is non-null. It follows that $P[I_s \notin \mathcal{A}_0]$ goes to one.

The statement of the theorem follows by combining the three cases.

Proof of Theorem 4

From (14), the log-likelihood is given by

$$\ell(\boldsymbol{p}, \boldsymbol{b}) = \sum_{i=1}^{n} \left(\ln \left(\sum_{j=1}^{m} \beta_{ij} q_j \right) \right) \tag{33}$$

Consider maximization of $\ell(\boldsymbol{p}, \boldsymbol{b})$ periodically with respect to $\boldsymbol{p}$ and $\boldsymbol{b}$. Given $(\boldsymbol{p}^{(n)}, \boldsymbol{b}^{(n)})$, the iterate at the nth stage, define the next iterate $(\boldsymbol{p}^{(n+1)}, \boldsymbol{b}^{(n+1)})$ by

$$\boldsymbol{b}^{(n+1)} = \begin{cases} \boldsymbol{b}^{(n)} & \text{if } n \text{ is even,} \\ \operatorname{argmax}\, \ell(\boldsymbol{p}^{(n)}, \boldsymbol{b}) & \text{if } n \text{ is odd,} \\ \boldsymbol{b} \in S_2 & \end{cases}$$

$$\boldsymbol{p}^{(n+1)} = \begin{cases} \boldsymbol{p}^{(n)} & \text{if } n \text{ is odd,} \\ \operatorname{argmax}\, \ell(\boldsymbol{p}, \boldsymbol{b}^{(n)}) & \text{if } n \text{ is even,} \\ \boldsymbol{p} \in S_1 & \end{cases} \tag{34}$$

where $S_1 = \{\boldsymbol{p} : \sum_{j=1}^{m} q_j = 1,\ 0 \le q_1, \ldots, q_m \le 1\}$ and $S_2 = \{\boldsymbol{b} : 0 \le b_1 \le \ldots \le b_k \le 1\}$. We shall show that the functions $\ell(\boldsymbol{p}, \cdot)$ and $\ell(\cdot, \boldsymbol{b})$ are concave over the convex sets S_1 and S_2, respectively, so that there exists a maximum at each iteration. Thus, in each stage there is an increase in the likelihood (14), which is bounded by $(km)^n$, and the sequence of partially maximized likelihoods converges. Under the conditions stated in the theorem, we shall also show that the objective function is strictly concave, so that the maximum at each stage is unique, with probability tending one as n_2 goes to infinity. Finally, since $S_1 \times S_2$ is a closed set, the sequence of maxima obtained at successive stages converges to a unique limit, with probability tending to one.

Let $\boldsymbol{B}$ be an $n \times m$ matrix with β_{ij} in the ijth position. For fixed $\boldsymbol{b}$, the partial derivative of (33) with respect to $\boldsymbol{p}$ is

$$\frac{\partial \ell}{\partial \boldsymbol{p}} = \sum_{i=1}^{n} \frac{B_i}{{B_i}^T \boldsymbol{p}}$$

where B_i is the ith row of $\boldsymbol{B}$ matrix. The second derivative or the Hessian is

$$\frac{\partial \ell}{\partial \boldsymbol{p} \partial \boldsymbol{p}^T} = -\sum_{i=1}^{n} \frac{B_i B_i^T}{({B_i}^T \boldsymbol{p})^2} \tag{35}$$

which is a non-positive definite matrix. Hence ℓ is a concave function over a convex and bounded domain, which ensures the existence of a maximum (Simon and Blume 1994). Now, we need to show that the probability of the Hessian matrix being negative definite goes to one. It is enough to show for any vector $\boldsymbol{u} \neq 0$,

$$P\left(\sum_{i=1}^{n} \frac{(B_i^T \boldsymbol{u})^2}{({B_i}^T \boldsymbol{p})^2} = 0\right) \to 0.$$

In other words, we need to show that for any arbitrary vector $\boldsymbol{u} \neq 0$,

$$P\left(B_i^T \boldsymbol{u} = 0 \quad \forall i\right) = P\left(\boldsymbol{B}\boldsymbol{u} = 0\right) \to 0. \tag{36}$$

It is clear from (15) that for an individual (say i) having exactly recalled age at landmark event, B_i has only one non-zero element. In this situation, the equation $B_i^T \boldsymbol{u} = 0$ implies that the corresponding element of $\boldsymbol{u}$ is zero. Further, Theorem 3 shows that, with probability tending to one, the columns of $\boldsymbol{B}$ correspond only to singleton members of $\mathcal{A}_0$ associated with individuals recalling age at event exactly. Therefore, with probability tending to one, the event $\boldsymbol{B}\boldsymbol{u} = 0$ coincides with the event $\boldsymbol{u} = 0$.

For fixed $\boldsymbol{p}$, the first derivative of (33) with respect to $\boldsymbol{b}$ is

$$\frac{\partial \ell}{\partial \boldsymbol{b}} = \sum_{i=1}^{n} \frac{\mathbf{A}_i \boldsymbol{p}}{{B_i}^T \boldsymbol{p}}$$

where $\mathbf{A}_i$ is the $k \times m$ matrix with the $(l, j)^{\text{th}}$ element given by $\frac{\partial \beta_{ij}}{\partial b_l}$.

The Hessian with respect to $\boldsymbol{b}$ is

$$\frac{\partial \ell}{\partial \boldsymbol{b} \partial \boldsymbol{b}^T} = -\sum_{i=1}^{n} \left(B_i^T \boldsymbol{p}\right)^{-2} \mathbf{A}_i \boldsymbol{p} \boldsymbol{p}^T \mathbf{A}_i^T \tag{37}$$

which is non-positive definite matrix. Hence ℓ is a concave function over a convex domain, it ensures the existence of a maximum (Simon and Blume 1994).

In order to prove the negative definiteness of the Hessian with probability tending to one, we need to show that for any arbitrary vector $\boldsymbol{v} \neq 0$,

$$P\left(\boldsymbol{v}^T \mathbf{A}_i \boldsymbol{p} = 0 \quad \forall i\right) \to 0. \tag{38}$$

From (15), it follows that for $i \in \mathcal{I}_2$,

$$\mathbf{A}_i \boldsymbol{p} = -\left(\sum_{j=1}^{m} q_j \cdot I(J_j \subset A_i)\right) \left(I(T_i \in A_{i1}), \ldots, I(T_i \in A_{ik})\right)^T, \tag{39}$$

which is a vector with a non-zero element exactly at one place. The condition $\boldsymbol{v}^T \mathbf{A}_i \boldsymbol{p} = 0$ is equivalent to the requirement that the element of $\boldsymbol{v}$ corresponding to the non-zero element of $\mathbf{A}_i \boldsymbol{p}$ is zero. On the other hand, as $n_2 \to \infty$,

$$\begin{aligned} &P\Big(\sum_{i \in \mathcal{I}_2} I\big((S_i - T_i) \in [x_l, x_{l+1}]\big) = 0\Big) \\ &\quad = \Big[P\Big((S_i - T_i) \in [x_l, x_{l+1}] | \delta_i \varepsilon_i = 1\Big)\Big]^{n_2} \to 0 \quad \forall l. \end{aligned}$$

Thus, for all $l = 1, \ldots, k$, there is at least one $i \in \mathcal{I}_2$ such that $T_i \in A_{il}$, with probability tending to one. Therefore, the condition $\boldsymbol{v}^T \mathbf{A}_i \boldsymbol{p} = 0 \quad \forall i \in \mathcal{I}_2$ reduces, with probability tending to one, to the requirement that all the elements of $\boldsymbol{v}$ are zero. Therefore, for $\boldsymbol{v} \neq 0$, we have $P\left(\boldsymbol{v}^T \mathbf{A}_i \boldsymbol{p} = 0, \quad \forall i\right) \leq P\left(\boldsymbol{v}^T \mathbf{A}_i \boldsymbol{p} = 0, \quad \forall i \in \mathcal{I}_2\right) \to 0$. Thus, the probability that the Hessian matrix defined in (37) is negative definite goes to one. This completes the proof.

Proof of Theorem 6

The proof relies on an application of Theorem 3.1 of Wang (1985), in the manner it was used by Gentleman and Geyer (1994). The said theorem makes use of five assumptions.

The first assumption requires a separable compactification of parameter space Θ. In the present case, the set $\overline{\Theta}$ serves this purpose. The Lévy distance can be used as metric, and the compactness follows by the Helley selection theorem. Homeomorphic mapping of $[t_{min}, t_{max}]$ to $[0, 1]$ can be used to establish separability (Billingsley 1968, p. 239). The equivalence class $\mathcal{E}$ defined by (24) is regarded as a single point in Θ. This takes care of the issue of non-identifiability as in Redner (1981).

Let, for $r = 1, 2, \ldots$, $V_r(F)$ be the Lévy neighborhood of $F \in \Theta$ with radius $1/r$. For such a sequence of decreasing open neighborhoods, Wang (1985)'s second assumption requires that, for any F_0 in Θ, there is a function $F_r : \overline{\Theta} \to V_r(F_0)$ such that (a) $\ell(F) - \ell(F_r(F))$ is locally dominated on $\overline{\Theta}$ and (b) $F_r(F)$ is in Θ if $F \in \Theta$. We define $F_r(F) = \frac{1}{r+1} F + \frac{r}{r+1} F_0$. Since $\|F_r(F) - F_0\| = \frac{1}{r+1} \|F - F_0\|$, and the Lévy distance is dominated by the Kolmogorov-Smirnov distance, it is clear that $F_r(F) \in V_r(F_0)$. Condition (b) is obviously satisfied. As for condition (a), note that

$$\begin{aligned}
&\sup_{F\in\overline{\Theta}} \left[\ell(F) - \ell(F_{F,r})\right] \\
&\quad = \sup_{F\in\overline{\Theta}} \ln \frac{\sum_{j=1}^{p} \alpha_{ij} \left(F(t_j) - F(t_{j}-)\right)}{\frac{1}{r+1}\left[\sum_{j=1}^{p} \alpha_{ij} \left(F(t_j) - F(t_{j}-)\right)\right] + \frac{r}{r+1}\left[\sum_{j=1}^{p} \alpha_{ij} \left(F_0(t_j) - F_0(t_{j}-)\right)\right]} \\
&\quad \le \ln(r+1),
\end{aligned}$$

which has finite expectation. Thus, $\ell(F) - \ell(F_r(F))$ is globally dominated on $\overline{\Theta}$.

The third assumption requires that $E[\ell(F) - \ell(F_r(F))] < 0$ for $F_0 \in \Theta$, $F \in \overline{\Theta}$, $F \neq F_0$. Here, F_0 needs to be interpreted as $\mathcal{E}$, and the result follows along the lines of the proof of Lemma 4.4 of Wang (1985).

The fourth and fifth assumptions require that $\ell(F) - \ell(F_r(F))$ is lower and upper semicontinuous for $F \in \overline{\Theta}$ except for a null set of points (which may depend on F only in the case of upper semicontinuity). Both the conditions follow from the portmanteau theorem (Billingsley 1968, p. 11), as argued by Gentleman and Geyer (1994). No null set needs to be invoked.

Since all the assumptions hold, the stated result follows from Theorem 3.1 of Wang (1985).

Proof of Theorem 7

Theorem 6 says that the Lévy distance of $\{\tilde{F}_n\}$ from the equivalence class $\mathcal{E}$ goes to zero almost surely as n goes to infinity, that is,

$$\inf_{F\in\mathcal{E}} d_L(\tilde{F}_n, F) \to 0 \qquad \text{as } n \to \infty \qquad \text{with probability 1.}$$

It follows that $P(\inf_{F\in\mathcal{E}} d_L(\tilde{F}_n, F) > \epsilon) \to 0$.

Using the fact that $P(\omega : \tilde{F}_n(\omega) = \hat{F}_n(\omega)) \to 1$, we conclude

$$P\left(\inf_{F\in\mathcal{E}} d_L(\hat{F}_n, F) > \epsilon\right) \to 0.$$

which proves the statement.

Proof of Theorem 8

Note that the equivalence class defined in (24) is the class of all distribution functions that have Kullback-Liebler ‘distance’ zero from the true unknown distribution. Let H be the probability measure corresponding to the density h, (which is determined by g, π_η and F through (6)). Let H_0 be the ‘true’ value of H. The Kullback-Liebler ‘distance’ between H and H_0 is defined as $D(H\|H_0) = \mu(h\log(\frac{h}{h_0}))$. By Jensen’s inequality it is easy to see that $D(H\|H_0) \geq 0$. The equality in Jensen’s inequality holds if and only if the argument of the log function is a constant, i.e.,

$$D(H\|H_0) = 0 \quad \text{iff} \quad H = H_0. \tag{40}$$

Under the conditions given in part (b) or (c) of Theorem 1, H completely identifies F. Hence, $H = H_0$ implies $F = F_0$. It follows that the true distribution of the time-to-event, F_0, is the only member of the equivalence class $\mathcal{E}$.

References

Aksglaede L, Sorensen K, Petersen JH, Skakkebak NE, Juul A (2009) Recent decline in age at breast development: The copenhagen puberty study. Pediatrics 123(5):932–939

Allison PD (1982) Discrete-time methods for the analysis of event histories. Sociol Methodol 13:61–98

Ayatollahi SM, Dowlatabadi E, Ayatollahi SA (2002) Age at menarche in iran. Ann Hum Biol 29(4):355–362

Beckett M, DaVanzo J, Sastry N, Panis C, Peterson C (2001) The quality of retrospective data: An examination of long-term recall in a developing country. J Hum Resour 36(3):593–625

Bergsten-Brucefors A (1976) A note on the accuracy of recalled age at menarche. Ann Hum Biol 3:71–73

Bickel PJ, Gotze F, van Zwet WR (1997) Resampling fewer than n observations: gains, losses, and remedies for losses. Stat. Sinica 7(1):1–31 Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995)

Bickel PJ, Sakov A (2008) On the choice of m in the m out of n bootstrap and confidence bounds for extrema. Stat. Sinica 18(3):967–985

Billingsley P (1968) Convergence of probability measures. Wiley, New York-London-Sydney

Cameron N (2002) Human growth and development. Academic Press, San Diego

Chumlea WC, Schubert CM, Roche AF, Kulin HE, Lee PA, Himes JH, Sun SS (2003) Age at menarche and racial comparisons in us girls. Pediatrics 11(1):110–113

Demirjian A, Goldstien H, Tanner JM (1973) A new system of dental age assessment. Ann Hum Biol 45:211–227

Efron B (1967) The two sample problem with censored data. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp 831–853

Eveleth PB, Tanner JM (1990) Worldwide variation in human growth, 2nd edn. Cambridge University Press, Cambridge

Gentleman R, Geyer CJ (1994) Maximum likelihood for interval censored data: consistency and computation. Biometrika 81(3):618–623

Hediger ML, Stine RA (1987) Age at menarche based on recall data. Ann Hum Biol 14:133–142

Hosmer DW, Lemeshow S (1999) Applied survival analysis: regression modeling of time to event data. Wiley, New York

ISI (2012) Annual report of the Indian Statistical Institute. http://library.isical.ac.in/jspui/handle/10263/5345?mode=full

Kalbfleisch JD, Lawless JF (1989) Inference based on retrospective ascertainment: an analysis of the data on transfusion-related aids. J Am Stat Assoc 84:360–372

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Keiding N, Begtrup K, Scheike TH, Hasibeder G (1996) Estimation from current-status data in continuous time. Lifetime Data Anal 2(2):119–129

Korn EL, Graubard BI, Midthune D (1997) Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. Am J Epidemiol 145:72–80

LeClere MJ (2005) Preface modeling time to event: applications of survival analysis in accounting, economics and finance. Rev Acc Financ 4:5–12

McKay HA, Bailey DB, Mirwald RL, Davison KS, Faulkner RA (1998) Peak bone mineral accrual and age at menarche in adolescent girls: A 6-year longitudinal study. J Pediatr 13:682–687

Mirzaei SS, Das R, Sengupta D (2015) Parametric estimation of menarcheal age distribution based on recall data. Scand J Stat. doi:10.1111/sjos.12107

Mirzaei SS, Sengupta D (2013) Nonparametric estimation of time-to-event distribution based on recall data in observational studies. Thecnical Report No. ASD/2013/7, Applied Statistical Unit, Indian Statistical Institue **7**. http://www.isical.ac.in/~asu/TR/TechRepASU201307.pdf

Rabe-Hesketh S, Yang S, Pickles A (2001) Multilevel models for censored and latent responses. Stat Methods Med Res 10(6):409–427

Redner R (1981) Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. Ann Stat 9(1):225–228

Sen B, Banerjee M, Woodroofe M (2010) Inconsistency of bootstrap: the Grenander estimator. Ann Stat 38(4):1953–1977

Simon CP, Blume L (1994) Mathematics for economists. W W Norton, New York

Sun J (2006) The statistical analysis of interval-censored failure time data. Springer, New York

Turnbull BW (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. J R Stat Soc Ser B 38:290–295

Wang JL (1985) Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics. Ann. Statist. 13(3):932–946