26

THE EMPIRICAL DISTRIBUTION FUNCTION WITH INCOMPLETE DATA

Alexander Tsodikov

Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Liebigstr. 27, D-04103 leipzig, Germany

ABSTRACT

This paper is concerned with the nonparametric maximum likelihood estimation (NPMLE) of a survivor function G from incomplete samples. An approach is suggested which allows for an exact maximization of the likelihood with finite number of steps by dynamic programming. The method provides in particular the maximum likelihood estimates (MLE) of the number and location of the support points of the empirical distribution. It offers an isotonic solution when the monotony property is not inherent to the empirical distribution function. Examples for doubly censored and discrete surveillance data are given.

Key words: Empirical distribution function; Dynamic programming algorithm; Nonparametric maximum likelihood estimation; Isotonic estimation; Discrete surveillance data; Doubly censored data

1 INTRODUCTION

This paper deals with nonparametric maximum likelihood estimation (NPMLE) of the survivor function of a nonnegative random variable which is incompletely observed. Incomplete data samples usually arise in the reliability analysis and biological studies when a failure or disease onset is unobservable either due to its nature or to a specific design of a study. It is not uncommon that a sample does not contain accurate times to failure, or there are quite few of them. In such cases the extent of information in a sample turns to be quite poor even if the sample itself is large. The estimate also termed the empirical survivor function is thought of as a nonincreasing function, which maximizes the "probability" to observe a given sample. The problem of nonparametric estimation of the distribution function from incomplete samples has been studied by many authors. For particular study designs as for example the follow-up with right noninformative censoring the problem is well-studied (Cox and Oakes(1983)). Further attempts to allow for other designs of a study were based on the "self-consistency" condition first formulated by Efron (1967) with respect to the product-limit estimate. The concept of "self-consistency" was extended by Turnbull (1976) who developed an iterative algorithm converging to the maximum likelihood estimate of the distribution function with arbitrary grouped, censored and truncated data. The algorithms based on the self consistency property although derived on the statistical basis are essentially close to the conventional nonlinear programming

algorithms. They are approximate iterative procedures requiring the initial point in the domain of local convergency, which is in most cases uncertain.

In the present paper an alternative approach is indicated, which allows for an exact solution with finite number of steps. However, an analytic effort is required to construct an algorithm specific to a given type of data. Proceeding from the Bellman's equations a particular algorithm is developed in detail. It was applied to either right or left censored data (termed doubly censored for short) and to data arising from a discrete surveillance study.

2 NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATE

Let the set $\mathcal{T} = \{t_i\}_{i=1}^m$ represent the observed ordered sample times from a homogeneous population with survivor function $G(t)$. Generally, the sample times are labelled to indicate the number and the type of events observed at t_i. For example, if right censored data are considered, each t_i is accompanied by the number of failures m_i and the number of censored items n_i observed at t_i. Given the data, the loglikelihood ℓ can be considered on the space of survivor functions $\mathcal{F}$. The problem of NPMLE will be to maximize ℓ on $\mathcal{F}$

$$\max_{G \in \mathcal{F}} \ell(G). \tag{26.1}$$

Usually ℓ is maximized by taking G a discrete survivor function with support points τ_j at some of t_i, and we will proceed from this assumption, which can be verified given specific likelihood. With the right censored data the support points will be at those t_i, for which $m_i > 0$ (the Kaplan-Meier estimate). Formally, the empirical survivor function will be defined as a right continuous step function with the vector of support points $s_n = (\tau_1, \ldots, \tau_n)$, $\tau_i \in \mathcal{T}$, $i = 1, \ldots, n$, $n \leq m$, $\tau_{n+1} \stackrel{\text{def}}{=} \tau_n$, $\tau_0 \stackrel{\text{def}}{=} 0$ and with values G_i of $G(t)$ in the intervals $[\tau_i, \tau_{i+1})$, $i = 0, \ldots, n$, $G_0 \stackrel{\text{def}}{=} 1$.

The above approach to derive an empirical distribution function is heuristic in a sense that the set $\mathcal{T}$ is random, although treated as fixed, and the number of parameters G_i grows with increase of the sample information (Kalbfleisch and Prentice (1980)). A rigorous study would require the estimate to be treated as a counting process in t (Fleming and Harrington (1991)), which is feasible only for simple estimates that are available in a closed form. It should be noted that if the set $\mathcal{T}$ is fixed in advance (as the case with discrete surveillance), standard likelihood theory applies.

Two equivalent definitions of $\mathcal{F}$ can be suggested

I $G(t)$ has a step at each t_i, $i = 1, \ldots, m$, and the step-values $\Delta G_i \stackrel{\text{def}}{=} G_{i-1} - G_i$ are nonnegative ($\Delta G_i = 0$ is allowed).

II $G(t)$ has steps at $\tau_i \in \mathcal{T}$, $i = 1, \ldots, n$, $n \leq m$, and the step-values are strictly positive $\Delta G_i > 0$, $i = 1, \ldots, n$. In other words $\mathcal{F} = \cup_{i=1}^{m} \mathcal{F}_i$, where $\mathcal{F}_i$ is a class of functions having exactly i strictly positive steps on $\mathcal{T}$.

According to the first definition we have to maximize ℓ with respect to G_i under the constraints $\Delta G_i \geq 0$, $i = 1, \ldots, m$. In what follows we will make use of the second definition to avoid the constrained maximization.

With complete or right censored data a conventional approach is to set s_n equal to the times of the observed failures and maximize the likelihood with respect to the remaining parameters G_i, which leads to well known estimates. With other types of data, postulating s_n may lead to an estimate which does not satisfy the monotony constraints as demonstrated in the next section. This is an indicator that the search for the optimal n, s_n should be performed.

3 SURPRISES OF INCOMPLETE DATA

3.1 Doubly censored data

Suppose that we have a sample of either right or left noninformatively censored data $\{t_i, \delta_i\}$, where

$$\delta_i = \begin{cases} 1, & \text{if } t_i \text{ is left censored} \\ 0, & \text{if } t_i \text{ is right censored,} \end{cases}$$

$i = 1, ..., m$. For such data

$$\ell = \sum_{i=1}^{n-1} m_i \ln(1 - G_i) + n_i \ln(G_i),$$

where

$$m_i = \sum_{k:\, t_k \in [\tau_i; \tau_{i+1})} \delta_k,$$

is the number of left censored cases in the ith interval between the adjacent support points, and

$$n_i = \sum_{k:\, t_k \in [\tau_i; \tau_{i+1})} (1 - \delta_k),$$

is the number of right censored cases in this interval. $i = 1, ..., n-2$, $\tau_n = t_m$,

$$m_{n-1} = \sum_{k:\, t_k \in [\tau_{n-1}; t_m]} \delta_k,$$

$$n_{n-1} = \sum_{k:\, t_k \in [\tau_{n-1}; t_m]} (1 - \delta_k).$$

Maximizing ℓ and ignoring the monotony constraint, we get the following estimate

$$\tilde{G}_i = \frac{n_i}{m_i + n_i}, \quad i = 1, ..., n-1 \tag{26.2}$$

Suppose we try to find $\tilde{G}$ with steps at each sample point $\tau_i = t_i$, $i = 1, ..., m$. If the data are untied, the estimate $\tilde{G}$ would look like a chaotic sequence of jumps from zero to one and back since in this case $\tilde{G}_i = 1 - \delta_i$ (Figure 1A).

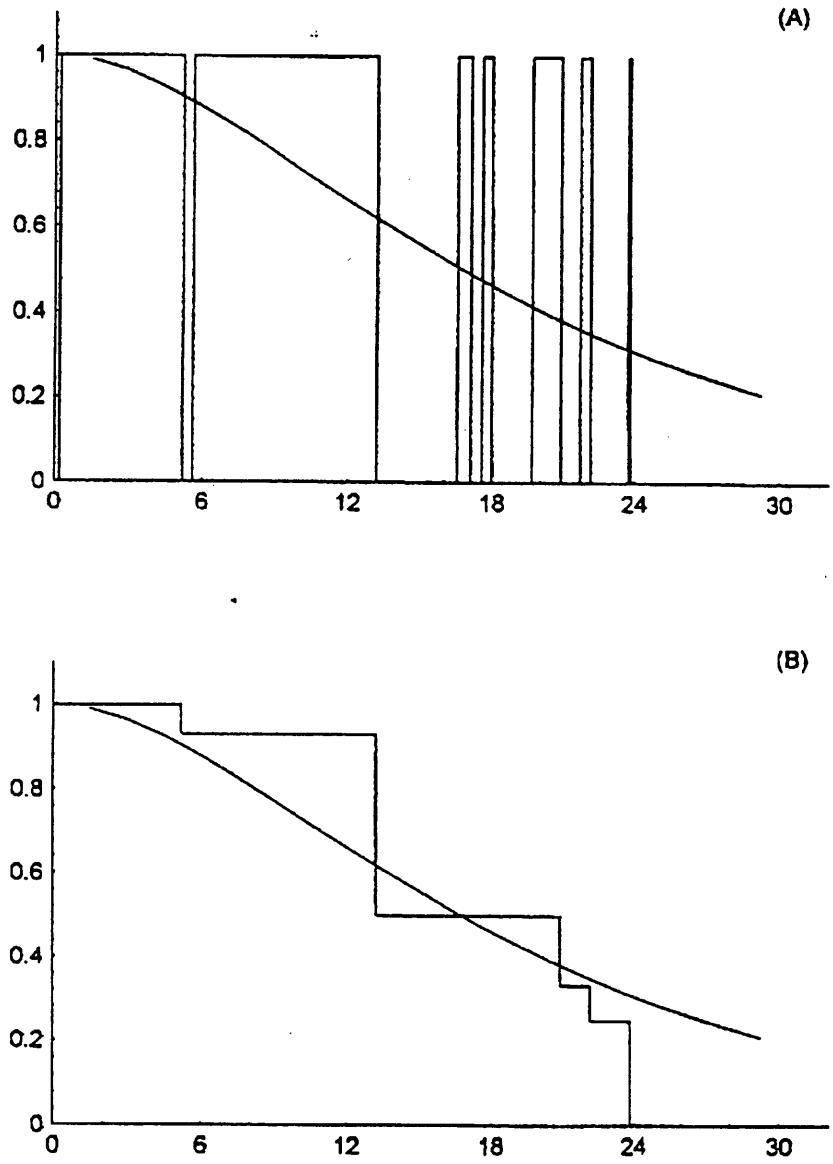

Figure 1: Survivor curves for doubly censored data. Solid line is the "true" survivor function used in computer simulations; stepwise curves are the nonparametric estimates $\tilde{G}$ (A) and the estimates $\hat{\tilde{G}}$ resulting from application of the dynamic programming algorithm (B).

This problem was addressed as early as in 1955 by Ayer et al. The estimate was considered as a function having exactly m steps and the step-probabilities were allowed to be zero (definition I of the class $\mathcal{F}$). In Ayer et al. (1955) an effective algorithm was found to solve the problem in the context of constrained optimization for this particular case. The algorithm was named for the pooled-adjacent-violators one and it has prompted the development of isotonic theory (Barlow et al. (1972)). According to this algorithm, the estimate $\tilde{G}_i = 1 - \delta_i$ is a starting point. The following step is repeated in arbitrary order. If some adjacent values G_i, G_{i+1} violate the monotony property, they are pooled i.e. the step-probability at the point i is set to zero and both G_i and G_{i+1} are substituted by a new G_i, the subsequent estimates being newly enumerated starting from $i+1$. It was shown by Ayer et al. (1955) that the resultant estimate is consistent and moreover is closer in average to the true survivor function than $\tilde{G}_i$. It is remarkable that the solution can be interpreted as the search for optimal grouping s_n (see Asselain et al. for more details).

With doubly censored data, application of the algorithm suggested in Section 4 provides the same solution as the algorithm by Ayer et al.

3.2 Discrete surveillance data

Suppose that a failure can be detected only by means of some test with probability p and that such tests are performed at times $\{t_i\}_{i=1}^m$. In addition the time to detection may be right censored. Such design of a study arises for example when a population of initially healthy individuals is repeatedly tested to detect cancers (cancer screening or surveillance) or in reliability theory when some units are tested to detect unobservable failures which cause damage. The times $\{t_i\}$ form a so-called surveillance strategy and are fixed in advance. There are sorts of control problems which can be solved to optimize a surveillance strategy. We refer the reader to Beichelt and Franken (1983), Parmigiani (1993), Tsodikov and Yakovlev (1991), Tsodikov (1992) for such examples and focus on the statistical aspect of discrete surveillance.

The sample generated by a discrete surveillance study consists of

N - the initial size of the target population;

m_i - the number of failures detected at the test performed at t_i;

n_i - the number of right censored observations in the interval $[t_{i-1}, t_i)$.

With $p = 1$ this design turns to be equivalent to that related to the life-table estimate. If in addition each individual is tested only once and t_i is the time of examination of the i-th individual, the design will be reduced to the doubly censored case.

In what follows we assume that $p < 1$ and t_i is the time when the whole population is tested all-at-once and that we have a sequence of such tests $i = 1, ..., m$. If a failure occurs

in some interval $[t_{i-1}, t_i)$ it will be detected at t_i with probability p, at t_{i+1} with probability $(1-p)p$, ... etc. In other words, the time of detection conditional on the failure entering $[t_{i-1}, t_i)$ is given by $t_{i-1+\xi}$, where ξ is a random variable following the geometric scheme with parameter p, truncated by the last test t_m.

Introduce the unconditional survivor function Q of the time to detection of failure, which is related to the distribution of time to failure by the following recurrence relations

$$\Delta G_i = \frac{1}{p}(\Delta Q_i - q\Delta Q_{i-1}), \tag{26.3}$$

$$i = 1, ..., m,\ q = 1-p,\ \Delta()_i = |()_i - ()_{i-1}|.$$

For each t_k the number of patients at risk will be given by $N_k = N - \sum_{i=1}^{k}(m_i + n_i),\ k = 0, ..., m,\ \sum_1^0 = 0$. The likelihood function is based on the observed detection process and is identical to the multinomial likelihood arising in the context of right censored data

$$\ell = \sum_{i=1}^{m} m_i \ln(\Delta Q_i) + n_i \ln(Q_{i-1}) + N_m \ln(Q_m). \tag{26.4}$$

Following the derivation of the Kaplan-Meier estimate we try to find an estimate of the time to failure survivor function which has a step at each t_i ignoring the monotony constraint. Such an estimate was derived in Tsodikov et al. (1995). The reasoning was quite simple. Using the invariance property of the ML estimate it is possible to estimate the time to detection survivor function Q as a Kaplan-Meier estimate $\hat{Q}$

$$\hat{Q}_i = \prod_{k=1}^{i} \frac{N_k}{N_{k-1} - n_k}.$$

Then using equations (26.3) expressing G in terms of Q it is easy to extract the estimate $\tilde{G}$. The resultant estimate may be written in a closed form

$$\Delta\tilde{G}_i = \frac{\prod_{k=1}^{i-1} N_k}{p\prod_{k=1}^{i}(N_{k-1} - n_k)}\left[m_i - qm_{i-1} + \frac{qm_{i-1}n_i}{N_{i-1}}\right]. \tag{26.5}$$

Since the estimate $\Delta\tilde{G}_i$ may be represented by a linear combination of the life-table ones, it inherits the properties of the latter, the consistency among them.

It is interesting to note that the monotony property is likely to be violated by (26.5) if the probability of mistake q is large. Indeed, in order that $\Delta\tilde{G}_i$ be nonnegative, $i = 1, ..., m$ we have to demand that the expression in square brackets in (26.5) be nonnegative. This can be written as

$$m_i \geq qm_{i-1}\frac{N_{i-1} - n_i}{N_{i-1}},\ i = 1, ..., m. \tag{26.6}$$

and the point becomes clear. For $p = 1$ $(q = 0)$ the inequalities (26.6) hold automatically. With increase of q the amount of information in the sample $\{m_i, n_i\}$ generally decreases

and this causes violation of the monotony property (Figure 2A). If the number of detected cases decreases too much at some t_i as compared to the previous test, it becomes unlikely that the survivor function has a step at t_i, since this would add to the number of detected cases. Again, a search for the optimal distribution of steps on $\mathcal{T}$ is suggested.

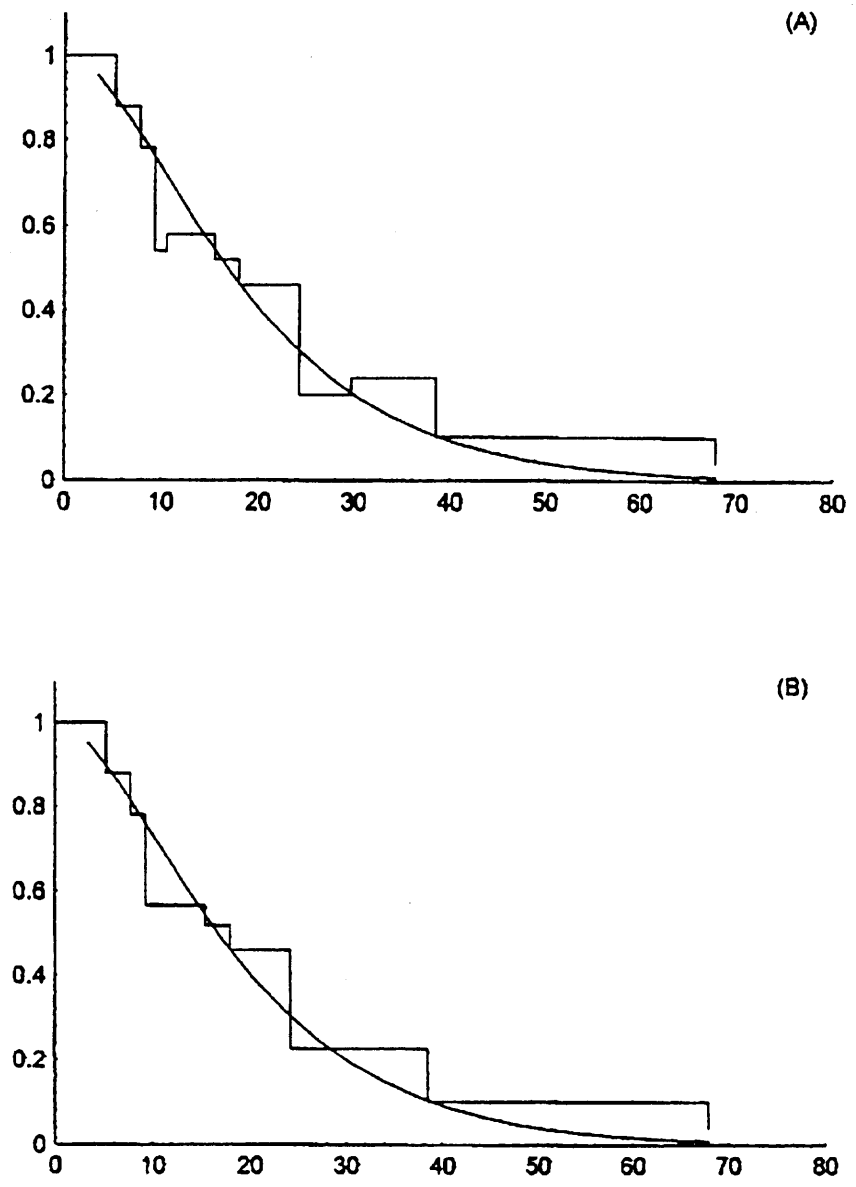

Figure 2: Survivor curves for discrete surveillance data. Solid line is the "true" survivor function used in computer simulations; stepwise curves are the nonparametric estimates $\tilde{G}$ (A) and the estimates $\hat{\tilde{G}}$ resulting from application of the dynamic programming algorithm (B).

4 THE ALGORITHM

Assume that given the vector of support points s_k and ignoring the constraints $\Delta G_i \geq 0$, it is easy to maximize the likelihood with respect to the remaining parameters G_i

$$\tilde{\ell}(s_n) = \max_{G_1,\ldots,G_n} \ell(G|s_n) = \ell(\tilde{G}|s_n) \tag{26.7}$$

by solving the likelihood equations $\frac{\partial \ell}{\partial G_i} = 0$, $i = 1,\ldots,n$. Instead of treating (26.1) as a problem with constraints, we reduce it to a number of unconstrained problems (26.7) and make use of the dynamic programming to improve the efficiency. The main idea will be to solve (26.7) for various n, s_n and to choose the monotone function with the largest likelihood. The procedure may be also looked upon as the search for a discrete model on $\mathcal{T}$ which would provide the best likelihood of the observed data.

Take some survivor function $G \in \mathcal{F}_k$ with k steps. Note that as some of ΔG_i tend to zero, the limit survivor function has less than k positive steps. For this reason the set $\cup_{i=1}^{k-1}\mathcal{F}_i$ will be the limit set of the set $\mathcal{F}_k$, $k = 2,\ldots,m$, which we denote by

$$\cup_{i=1}^{k-1}\mathcal{F}_i = \Gamma(\mathcal{F}_k),\ k = 2,\ldots,m. \tag{26.8}$$

Consequently, the closing $\bar{\mathcal{F}}_k$ of the set $\mathcal{F}_k$ contains all functions that have not more than k positive steps $\bar{\mathcal{F}}_k = \mathcal{F}_k \cup \Gamma(\mathcal{F}_k) = \cup_{i=1}^{k}\mathcal{F}_i$. Evidently, the sets $\bar{\mathcal{F}}_k$ are nested

$$\bar{\mathcal{F}}_1 \subset \bar{\mathcal{F}}_2 \subset \ldots \subset \bar{\mathcal{F}}_m = \mathcal{F}. \tag{26.9}$$

Since the solution of (26.1) $\hat{G}$ is a function which has some positive steps, it should be a solution to one of the problems

$$\sup_{G\in\mathcal{F}_n} \ell(G),\ n = 1,\ldots,m, \tag{26.10}$$

with sup attained in an inner point of $\mathcal{F}_n$, which means it must be the solution of (26.7) and must satisfy the likelihood equations and the monotony constraints for some unknown n, s_n (to be found). If the solution of the likelihood equations (given n, s_n) does not satisfy the monotony constraints, then $\max_{1>G_1>\ldots>G_n>0} \ell(G|s_n)$ is attained at the limit point of the set $\mathcal{F}_n$ (which does not belong to $\mathcal{F}_n$). Consequently, we can replace (26.10) by

$$\max_{s_n:\tilde{G}\in\mathcal{F}_k} \tilde{\ell}(s_n),\ \tilde{\ell}(s_n) = \ell(\tilde{G}|s_n). \tag{26.11}$$

Finally, we arrive at the problem

$$\max_{n=1,\ldots,m} \max_{s_n:\tilde{G}\ -\ \text{monotone}} \tilde{\ell}(s_n). \tag{26.12}$$

By virtue of (26.9) the optimal n will be the largest possible one such that the monotone solution of (26.7) exists for some s_n (the estimate must be as detailed as possible). If by chance the solution of (26.7) with $\tilde{G}$ having a step at each t_i is monotone, then it is the optimal one.

Since the step-times comprising s_n are taken from the observed sample $\mathcal{T}$, we may represent s_n by the indexes (ranks) λ_j, $j = 1, \ldots, n$ of those t_i at which a positive step of the survivor function is assumed $\tau_i = t_{\lambda_i}$, $i = 1, \ldots, n$, $\lambda_0 \stackrel{\text{def}}{=} 0$, $\lambda_n \stackrel{\text{def}}{=} m$. So, to find $\max\limits_{s_n}$ we could look over $\binom{m-1}{n-1}$ combinations of $\{\lambda_i\}_{i=1}^{m-1}$ on the grid $\{1, \ldots, m-1\}$ and compute the likelihood $\tilde{\ell}(s_n)$ for each of them. However this would be an inefficient procedure. Further improvements in the efficiency have to do with the specific properties of the likelihood and the dynamic programming technique.

In the examples considered in Sections 3,5 the likeihood $\tilde{\ell}(s_n)$ can be represented in the form

$$\tilde{\ell}(s_n) = \sum_{i=1}^{n-1} \varphi(\lambda_i, \lambda_{i+1}), \tag{26.13}$$

where φ is some function of the adjacent step-times. The monotony constraints are supposed to be in the form

$$\psi_i = \psi(\lambda_{i-1}, \lambda_i, \lambda_{i+1}) > 0, \quad i = 2, ..., n-1, \ \psi_n(\lambda_{n-1}, \lambda_n) > 0.$$

The efficiency of the algorithm depends on the "depth of interaction", i.e. on how many variables appear to be the arguments of the functions φ and ψ. For a particular example the main concern will be to derive specific expressions for φ and ψ.

The next step will be to derive the Bellman's recursive equations for the problem (26.12) with $\tilde{\ell}$ given by (26.13). Imagine that the optimal step-time $\tau_{n-1}(\lambda_{n-1})$ is available. Then the following problem remains to be solved

$$\max_{\substack{\lambda_1, ..., \lambda_{n-2} \\ \psi_2 > 0, ..., \psi_{n-1} > 0}} \sum_{i=1}^{n-2} \varphi(\lambda_i, \lambda_{i+1}) \tag{26.14}$$

Reasoning from this observation we may write

$$\max_{s_n, \psi_i > 0, i=2,...,n} \ell(s_n) = \max_{\lambda_{n-1}, \psi_n > 0} \ell^{(n-2)}(\lambda_{n-1}, \lambda_n) + \varphi(\lambda_{n-1}, \lambda_n), \tag{26.15}$$

where $\ell^{(n-2)}(\lambda_{n-1}, \lambda_n) = \sum\limits_{i=1}^{n-2} \varphi(\lambda_i, \lambda_{i+1})$ given that $\lambda_1, ..., \lambda_{n-2}$ are the solutions of (26.14).

Along the similar lines we get

$$\max_{\substack{\lambda_1, ..., \lambda_{k-1} \\ \psi_2 > 0, ..., \psi_k > 0}} \sum_{i=1}^{k-1} \varphi(\lambda_i, \lambda_{i+1}) = \max_{\lambda_{k-1}, \psi_k > 0} \ell^{(k-2)}(\lambda_{k-1}, \lambda_k) + \varphi(\lambda_{k-1}, \lambda_k), \tag{26.16}$$

where $\ell^{(k-2)}(\lambda_{k-1}, \lambda_k) = \sum\limits_{i=1}^{k-2} \varphi(\lambda_i, \lambda_{i+1})$ given that $\lambda_1, ..., \lambda_{k-2}$ are the solutions of the problem

$$\max_{\substack{\lambda_1, ..., \lambda_{k-2} \\ \psi_2 > 0, ..., \psi_{k-1} > 0}} \sum_{i=1}^{k-2} \varphi(\lambda_i, \lambda_{i+1}), \ k = 3, 4, ..., n-1. \tag{26.17}$$

If the solution of (26.17) is empty $\ell^{(k-2)}$ is set to $-\infty$.

The recursive relations (26.15), (26.16) present specific forms of the Bellman's equations. Based on (26.15), (26.16) the k-th step of the procedure ($k = 2, ..., m-1$) is given by

- Step k.

  - For every $\lambda_{k-1}, \lambda_k : k-1 \le \lambda_{k-1} < \lambda_k \le m$ find

$$\max_{\substack{\lambda_{k-2} : k-2 \le \lambda_{k-2} < \lambda_{k-1} \\ \psi_{k-1} > 0}} \ell^{(k-3)}(\lambda_{k-2}, \lambda_{k-1}) + \varphi(\lambda_{k-2}, \lambda_{k-1}),$$

    If $k > 3$ then λ_i, $i = 1, ..., k-3$ and $\ell^{(k-3)}$ are known from step $k-1$, otherwise $\ell^{(k-3)} = 0$.

  - Register $\ell^{(k-2)}(\lambda_{k-1}, \lambda_k)$ and the solution $\lambda_1, ..., \lambda_{k-2}$ as functions (arrays) of λ_{k-1}, λ_k to be used at the next step.
  - Find

$$\max_{\substack{\lambda_{k-1} : k-1 \le \lambda_{k-1} < m \\ \psi_n(\lambda_{n-1}, m) > 0,\ n = k}} \ell^{(k-2)}(\lambda_{k-1}, m) + \varphi(\lambda_{k-1}, m).$$

    The solution of the above problem is equivalent to that of the problem

$$l_k = \max_{s_k : \tilde{G} - \text{monotone}} \ell(\tilde{G}|s_k).$$

In the above step maximization is performed by exhaustion. The combinations of variables which does not satisfy the constraint are skipped. If the solution appears to be empty, l_k is set to $-\infty$. At the first step we just compute $l_1 = \ell(\tilde{G})$, where $\tilde{G}$ has just one step at t_m. At the last step (m) we find $\ell_m = \ell(\tilde{G})$ with $\tilde{G}$ having steps at each t_i, $i = 1, \ldots, m$. Finally, $\tilde{G}$ corresponding to the largest l_k will give the sought-for NPMLE $\hat{G}$.

4.1 Complexity

As an elementary substep of the procedure we will take the computation of $\ell^{(j)} = \ell^{(j-1)} + \varphi$ and checking the constraint ψ_{j+1} performed as an elementary unit of the exhaustive search at each step of the procedure. To characterize the complexity we use the total number C of elementary substeps contained in the steps $1, ..., m$: $C = \sum_{k=1}^{m} C_k$, where C_k is the number of elementary substeps at step k. We have

$$\begin{aligned} C_1 &= 1, \\ C_2 &= \binom{m-1}{1} = m-1 \end{aligned}$$

$$\begin{aligned} C_k &= |\{\lambda_{k-2}, \lambda_{k-1}, \lambda_k : k-2 \le \lambda_{k-2} < \lambda_{k-1} < \lambda_k \le m\}| + \\ &\quad |\{\lambda_{k-1} : k-1 \le \lambda_{k-1} < m\}| = \\ &\quad \binom{m-k+3}{3} + m - k + 1, \quad k = 3, ..., m-1, \\ C_m &= 1. \end{aligned} \tag{26.18}$$

Summing up the C_k in (18) we get

$$C = \frac{m}{2}(m-1) + \sum_{i=3}^{m} \binom{i}{3}.$$

Let us find $S = \sum_{i=3}^{m} \binom{i}{3}$. Consider the function

$$F(s) = \sum_{j=3}^{m} (1+s)^j$$

and note that

$$S = \frac{1}{6} \frac{d^3 F(s)}{ds^3}|_{s=0} = \frac{1}{6} \left[\sum_{j=3}^{m} j^3 - 3 \sum_{j=3}^{m} j^2 + 2 \sum_{j=3}^{m} j \right]. \tag{26.19}$$

It can be proved by induction that

$$\sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6}, \quad \sum_{j=1}^{m} j^3 = \left(\frac{m(m+1)}{2} \right)^2. \tag{26.20}$$

Substituting (26.20) in (26.19) after a little algebra we obtain

$$S = \frac{m(m+1)}{24} \left[m^2 - 3m + 2 \right],$$

$$C = S + \frac{m}{2}(m-1) \sim \frac{m^4}{24}.$$

Practically the functions φ and ψ can be tabulated before the algorithm starts, so that each substep contain only a couple of elementary arithmetic operations.

A typical cancer surveillance program takes about 10 years with a testing frequency between "once half a year" and "once every two years". That means, we have to compute φ and ψ not more than 7000 times, a simple task for a PC.

5 THE ESTIMATE WITH DISCRETE SURVEILLANCE DATA

To apply the algorithm we need to generalize (26.5) to allow for the steps to occur arbitrary on the set $\mathcal{T} = \{t_i\}_{i=1}^{m}$. Because of a monotony restriction, the argument with the Kaplan-Meier estimate for the time to detection survivor function Q used in Subsection 3.2 breaks

down. To derive an estimate which generally has less than m steps, introduce an n-step-function P which coincides with (m-step-function) Q at the step-points t_{λ_i-1}, $i = 1, ..., n+1$, $\lambda_{n+1} \stackrel{\text{def}}{=} n+1$ such that

$$P(t_{\lambda_i} - 0) = Q(t_{\lambda_i} - 0), \quad i = 1, ..., n,$$

or

$$P_{i-1} = Q_{\lambda_i - 1}, \; i = 1, ..., n+1, \; P_0 = 1, \; P_i = P(\tau_i).$$

By definition the function P has as many steps as G and we can look for an argument similar to that of the Subsection 3.2. It is convenient to represent P_i in terms of variables

$$r_i = \frac{P_i}{P_{i-1}}, \; 0 \le r_i \le 1, \quad i = 1, ..., n.$$

It is not difficult to show that the likelihood (26.4) can be rewritten in the form (26.13) with

$$\varphi(\lambda_i, \lambda_{i+1}) = \left(N_{\lambda_{i+1}-1} - n_{\lambda_{i+1}}\right) \ln(r_i) + M_i^{(1)} \ln(1 - r_i) + M_i^{(3)} +$$

$$\sum_{k=\lambda_i}^{\lambda_{i+1}-1} n_{k+1} \ln\left[1 - (1 - r_i)\frac{1 - q^{k-\lambda_i} + 1}{1 - q^{\lambda_{i+1}-\lambda_i}}\right], \quad i = 1, ..., n-1, \tag{26.21}$$

where

$$M_i^{(1)} = \sum_{k=\lambda_i}^{\lambda_{i+1}-1} m_k,$$

$$M_i^{(2)} = \sum_{k=\lambda_i}^{\lambda_{i+1}-1} m_k (k - \lambda_i) \ln(q),$$

$$M_i^{(3)} = M_i^{(1)} \ln\left(\frac{1-q}{1 - q^{\lambda_{i+1}-\lambda_i}}\right) + M_i^{(2)}.$$

The derivation of (26.21) as well as other details are outlined in the Appendix. The function ψ which describes the monotony constraints is given by

$$\psi = \frac{1 - r_i}{1 - q^{\lambda_{i+1}-\lambda_i}} - \frac{1 - \frac{1}{r_{i-1}}}{1 - \frac{1}{q^{\lambda_i - \lambda_{i-1}}}}, \quad \psi_n = \frac{1 - r_n}{p} - \frac{1 - \frac{1}{r_{n-1}}}{1 - \frac{1}{q^{m-\lambda_{n-1}}}}, \tag{26.22}$$

$i = 1, ..., n-1$. And finally the variables $r_i = r_i(\lambda_i, \lambda_{i+1})$ are obtained by solving the algebraic equations

$$\frac{N_{\lambda_{i+1}-1} - n_{\lambda_{i+1}}}{r_i} + \sum_{k=\lambda_i}^{\lambda_{i+1}-1} \frac{n_{k+1}}{r_i + \frac{q^{k-\lambda_i+1} - q^{\lambda_{i+1}-\lambda_i}}{1 - q^{k-\lambda_i+1}}}} = \frac{M_i^{(1)}}{1 - r_i}, \tag{26.23}$$

$i = 1, ..., n$. It is pleasant to note that the right part of (26.23) is increasing with respect to r_i from $M_i^{(1)}$ ($r_i = 0$) to ∞ ($r_i = 1$), while the left part is decreasing from ∞ to its

value at $r_i = 1$. Therefore there exists the unique solution of (26.23) for each $i = 1, ..., n$. The function $\tilde{G}$ is given by

$$\Delta\tilde{G}_i = \prod_{k=1}^{i-1} r_k \left[\frac{1 - r_i}{1 - q^{\lambda_{i+1} - \lambda_i}} - \frac{1 - \frac{1}{r_{i-1}}}{1 - \frac{1}{q^{\lambda_i - \lambda_{i-1}}}} \right], \tag{26.24}$$

$$i = 1, ..., n, \; r_0 \stackrel{\text{def}}{=} 1, \; \Delta\tilde{G}_0 = 0, \; \prod_{1}^{0} = 1.$$

Once the algorithm has been applied and the values $r_i(\lambda_i, \lambda_{i+1})$ have been found, the sought-for estimate $\hat{G}$ is drawn from (26.24).

It is not difficult to verify that in case $\lambda_i = i, \; i = 1, ..., m, \; n = m$ we will have (26.5) instead of (26.24).

Another particular case is of interest. If the censoring is of type I (observations are censored only by the end of the study), i.e. if $n_i = 0, \; i = 1, ..., n$, then the roots of (26.23) are available in the closed form

$$r_j = \frac{N_{\lambda_{j+1}-1}}{N_{\lambda_j - 1}}, \quad j = 1, ..., n.$$

6 NUMERICAL EXAMPLE

For an example of application of the dynamic programming algorithm we have simulated a sample $\{x_i\}$ of 50 points from the two-parameter Gamma distribution with density

$$f(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}, \quad \alpha > 1, \quad t \geq 0 \, .$$

The shape parameter α and the scale parameter β were taken to be 2.0 and 0.1, respectively. Then the simulated observed process was imposed to generate the observed data for the doubly censored design and for the discrete surveillance one. The doubly censored times $\{t_i\}_{i=1}^m$, $m = 50$ were generated from the uniform distribution on the interval [0,30] and the i-th member of the initial sample was taken as right censored if $x_i > t_i$ and left censored otherwise. As to the discrete surveillance, the test times $\{t_i\}_{i=1}^m$, $m = 10$ were taken at each 5-th point of the initial sample. Then the detection process was organized. For each failure entering the interval $[t_{i-1}, t_i)$, $i = 1, ..., m$ the value of ξ was generated from geometric distribution (with p=0.5) truncated by $m - i + 1$. The time of detection was taken to be $t_{i-1+\xi}$. If undetected until t_m the observation was declared as censored by the end of the study. In Figures 1 and 2 the results of estimation of the empirical survivor function are shown for the doubly censored design and for the discrete surveillance one, respectively. The (A) part of both figures contains the true curve and the estimate $\tilde{G}$ having steps at each t_i disregarding the monotony constraint. The part (B) presents the true curve and the estimates of the empirical survivor function resulting from application

of the algorithm. From these figures it is evident that in both examples the estimate should be taken less detailed than the samples which are "noisy" and therefore unable to provide enough information to make the estimate $\tilde{G}$ monotone.

7 CONCLUSION

The proposed algorithm is primarily oriented to the analysis of small (or low information) samples. An indication of such a case could be either the violation of monotony by $\tilde{G}$ or its instability to the choice of grouping. With large samples with enough information for the asymptotic ML theory to be applied with respect to $\tilde{G}$, we do not need this algorithm, because in this case $\tilde{G}$ is monotone and stable as a consistent estimate and therefore can be made as detailed as desired. However, for example in medical applications this is not the case more often than not. The field of application of the proposed algorithm is not limited to searching for the empirical distribution function. It might be reasonable for example to minimize the chi-square statistics when testing a parametric hypothesis with respect to grouping n, s_n thus avoiding the problems associated with instability of the conventional test to the choice of grouping. However, the asymptotic distribution of the modified statistics may turn to be other than χ^2.

APPENDIX A

We are going to derive (26.21)-(26.24). In doing so we proceed from the likelihood in the form (26.4), where the time to detection distribution Q and the time to failure distribution G are linked by

$$\Delta Q_i = Q(t_{i-1}) - Q(t_i) = p(G(t_{i-1}) - G(t_i)) + q\Delta Q_{i-1}. \tag{A.1}$$

Since the function G actually has steps at points $\tau_i = t_{\lambda_i}$, $i = 1, ..., n$, we have

$$G(t_{j-1}) - G(t_j) = 0,\ j = \lambda_i, \lambda_i + 1, ..., \lambda_{i+1} - 1,\ i = 1, ..., n-1. \tag{A.2}$$

Recall the step-function P, introduced in Section 5

$$P_{i-1} = Q_{\lambda_i - 1},\ i = 1, ..., n+1,\ P_0 = 1,\ P_i = P(\tau_i),\ \lambda_{n+1} \stackrel{\text{def}}{=} m+1. \tag{A.3}$$

and its presentation in terms of variables r_i

$$P_i = \prod_{k=1}^{i} r_k, \quad i = 1, ..., n.$$

Using (A.1) and (A.2) the values $\Delta Q_{\lambda_i} = Q(\tau_i - 0) - Q(\tau_i)$ can be expressed in terms of the function P

$$\Delta Q_{\lambda_i} = \Delta P_i \frac{1-q}{1-q^{\lambda_{i+1}-\lambda_i}}, \quad i = 1, ..., n \tag{A.4}$$

Let us rewrite the likelihood (26.4)

$$\ell = \sum_{i=1}^{n} \sum_{k=\lambda_i}^{\lambda_{i+1}-1} [m_i \ln(\Delta Q_i) + n_{i+1} \ln(Q_i)]. \tag{A.5}$$

In (A.5) use was made of the following remark. For $i = 0, 1, ..., \lambda_1 - 1$ it must be $m_i = 0$, $\Delta Q_i = 0$, $Q_i = 1$. In other words the prevalence must be zero before the first step-point, since with probability 1 there are no failures on $[0, \tau_1)$. From (A.1) and (A.2) we also get

$$\Delta Q_k = \Delta Q_{\lambda_i} q^{k-\lambda_i}, \quad k = \lambda_i, \lambda_i + 1, ..., \lambda_{i+1} - 1, \tag{A.6}$$

$$i = 0, ..., n, \quad \lambda_0 \stackrel{\text{def}}{=} 0, \ \lambda_{n+1} \stackrel{\text{def}}{=} m + 1, \ \Delta Q_0 \stackrel{\text{def}}{=} 0.$$

Equation (A.6) and (A.5) combined give

$$\ell = \sum_{i=1}^{n} \Bigg[M_i^{(1)} \ln(\Delta Q_{\lambda_i}) + M_i^{(2)} + \sum_{k=\lambda_i}^{\lambda_{i+1}-1} n_{k+1} \ln\left(\bar{Q}_{\lambda_i - 1} - \Delta Q_{\lambda_i} \frac{1-q^{k-\lambda_i+1}}{1-q} \right) \Bigg], \tag{A.7}$$

$$M_i^{(1)} = \sum_{k=\lambda_i}^{\lambda_{i+1}-1} m_k, \quad M_i^{(2)} = \sum_{k=\lambda_i}^{\lambda_{i+1}-1} m_k (k - \lambda_i) \ln(q), \quad i = 1, ..., n.$$

With the help of (A.3) and (A.4) we rewrite (A.7) in terms of variables r_i

$$\ell = \sum_{i=1}^{n} \Bigg[\left(M_i^{(1)} + N_i^{(1)} \right) \sum_{j=1}^{i-1} \ln(r_j) + M_i^{(1)} \ln(1 - r_i) + M_i^{(3)} + \sum_{k=\lambda_i}^{\lambda_{i+1}-1} n_{k+1} \ln\left\{ 1 - (1 - r_i) \frac{1-q^{k-\lambda_i+1}}{1-q^{\lambda_{i+1}-\lambda_i}} \right\} \Bigg] \tag{A.8}$$

$$M_i^{(3)} = M_i^{(1)} \ln\left(\frac{1-q}{1-q^{\lambda_{i+1}-\lambda_i}} \right) + M_i^{(2)},$$

$$N_i^{(1)} = \sum_{k=\lambda_i}^{\lambda_{i+1}-1} n_{k+1}, \quad i = 1, ..., n, \ \sum_{1}^{0} = 0.$$

Next, changing the order of summation we obtain

$$\sum_{i=1}^{n} \left(M_i^{(1)} + N_i^{(1)} \right) \sum_{j=1}^{i-1} \ln(r_j) = \sum_{j=1}^{n-1} \ln(r_j) \sum_{i=j}^{n-1} \left(M_{i+1}^{(1)} + N_{i+1}^{(1)} \right) =$$

$$\sum_{j=1}^{n-1} \ln(r_j) \left(N_{\lambda_{j+1}-1} - n_{\lambda_{j+1}} \right).$$

Substituting this in (A.8) we arrive at the presentation (26.13) for the likelihood with φ in the form (26.21).

To maximize ℓ with respect to r_j we take the derivatives

$$\frac{\partial \ell}{\partial r_j} = \frac{\partial \varphi(\lambda_j, \lambda_{j+1})}{\partial r_j} =$$

$$\frac{N_{\lambda_{j+1}-1} - n_{\lambda_{j+1}}}{r_j} - \frac{M_j^{(1)}}{1-r_j} + \sum_{k=\lambda_j}^{\lambda_{j+1}-1} \frac{n_{k+1}}{r_j + \frac{q^{k-\lambda_j+1} - q^{\lambda_{j+1}-\lambda_j}}{1-q^{k-\lambda_j+1}}} = 0,$$

$j = 1, ..., n-1$. It is not difficult to note that this expression can be extended also to $j = n$ by virtue of the conventions mentioned above. In particular,

$$\frac{\partial \ell}{\partial r_n} = \frac{N_m}{r_n} - \frac{m_m}{1-r_n} = 0,$$

and consequently

$$r_n = \frac{N_m}{N_{m-1} - n_m}.$$

The latter expression as well as the function $\varphi(\lambda_n, \lambda_{n+1})$ does not depend on the sought-for $\{\lambda_i\}_{i=1}^{n-1}$ and therefore $\varphi(\lambda_n, \lambda_{n+1})$ can be omitted in ℓ. The representation of the monotony constraints

$$\Delta G_i > 0, \quad i = 1, ..., n$$

in terms of r_i follows directly from (26.24), the latter being due to (A.1), (A.4) and (A.6).

Acknowledgements

Research is supported by the grant Lo 342/6-1 of the Deutsche Forschungsgemeinschaft.

REFERENCES

[1] Asselain, B., Fourquet, A., Hoang, T., Tsodikov, A. and A. Yakovlev, (to appear): "A parametric regression analysis of breast cancer recurrence after the conservative treatment of primary tumor". *Statistics and Probability Letters.*

[2] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and E. Silverman, (1955): "An empirical distribution function for sampling with incomplete information". *Annals of Mathematical Statistics*, **26**, 641-647.

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

[4] Beichelt, F. and P. Franken, (1983): *Zuverlässigkeit und Instandhaltung.* VEB Verlag Technik, Berlin.

[5] Cox, D.R. and D. Oakes, (1983): *Analysis of survival data*, Chapman and Hall, London.

[6] Efron, B., (1967): "The two-sample problem with censored data". In: *Proceedings of the 5-th Berkeley Simposium in Mathematical Statistics IV*, (Prentice-Hall, New York, 831-853.

[7] Fleming, T.R. and D.P. Harrington, (1991): *Counting processes and survival analysis.* John Wiley and Sons, NY.

[8] Kalbfleisch, J.D., Prentice, R.L., (1980): *The statistical analysis of failure time data.*, John Wiley and Sons, NY.

[9] Parmigiani, G., (1993): "On optimal screening ages". *Journal of the American Statistical Association*, **88**, 622-628.

[10] Tsodikov, A.D., (1992): "Screening under uncertainty. Games approach". *Systems Analysis. Modeling and Simulation*, **9**, 259-262.

[11] Tsodikov, A.D., Asselain, B., Fourquet, A., Hoang, T. and A.Yu. Yakovlev, (1995): "Discrete strategies of cancer post-treatment surveillance. Estimation and optimization problems". *Biometrics*, **51**, 437-447.

[12] Tsodikov, A.D. and A.Yu. Yakovlev, (1991): "On the optimal policies of cancer screening". *Mathematical Biosciences*, **107**, 21-45.

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