Analysis of Censored Data
IMS Lecture Notes – Monograph Series (1995) Volume 27

Nonparametric estimators for interval censoring problems*

Piet Groeneboom
Delft University of Technology

Abstract

We study weighted least squares estimators for the distribution function of observations which are only visible via interval censoring, i.e., in the situation where one only has information about an interval to which the variable of interest belongs and where one cannot not observe it directly. The least squares estimators are shown to be closely related to nonparametric maximum likelihood estimators (NPMLE's) and to coincide with these in certain cases. New algorithms for computing the estimators are presented and it is shown that they converge from any starting point (in contrast with the EM-algorithm in this situation). Finally, the estimation of non-smooth and smooth functionals of the model is considered; for the latter case, we discuss $\sqrt{n}$-consistency and efficiency of the NPMLE.

1 Introduction

An extensive statistical theory exists for treating right censored data. Much less is known about more general types of censorship. This paper considers estimators for data subject to interval censoring. In this situation one only has information about an interval to which the observation of interest belongs; so only indirect information about the observation of interest is available.

Most of the time the interval will be a time interval, but the following interesting spatial version of this situation was brought to our attention by professor Dietz. In examinations of skin tissue, possibly affected by skin cancer, successive (roughly) circular incisions are made to determine the region of affected tissue; in this case one tries to estimate the smallest "safe" radius determining the region on which the operation should take place. On the one hand one tries to minimize the number of incisions, but on the

*AMS 1991 subject classifications. 60F17, 62E20, 62G05, 62G20, 45A05.
Key words and phrases. nonparametric maximum likelihood, empirical processes, asymptotic distributions, asymptotic efficiency, Fredholm integral equations.

other hand making too few incisions might result in an estimate which is too rough. Clearly statistical information about the estimates based on interval censoring could be very valuable here.

Aids research provides other important examples of interval censoring; usually the time of onset of a certain stage of the disease is unknown, but often indirect information about this is available.

In this paper we will concentrate on the following two cases of interval censoring:

Case 1. For each individual we make one observation and observe whether or not the event of interest has occurred before the time of observation. Such data arise for instance in cross-sectional studies.

Case 2. Two examinations at particular times are made so that it is known whether the event happened before the first observation (left censored), between the two observations (interval censored) or after the second observation (right censored).

AYER ET AL. (1955) derived the nonparametric maximum likelihood estimator (NPMLE) of the distribution function for Case 1 and proved that it is consistent. In this case the NPMLE can be calculated in a finite number of steps using the "pool adjacent violators" algorithm.

PETO (1973) considers the NPMLE for the more general Case 2. He suggests that pointwise standard errors for the survival curve can be estimated from the inverse of the Fisher information, which, however, is not correct.

Turnbull in TURNBULL (1974) and TURNBULL (1976) proposes the use of an EM algorithm to compute the NPMLE in interval censored problems. On the other hand, it is shown in GROENEBOOM AND WELLNER (1992), Chapter 1, Part II, that the "self-consistency" equation is a necessary but not a sufficient condition for the NPMLE. The EM-algorithm may therefore converge to some inconsistent estimator. Further, even if the starting function is such that the algorithm *will* converge to the NPMLE, the rate of convergence is generally very slow. Finally, the self-consistency equations have not been successful in developing distribution theory. For these reasons we turn to another approach, based on isotonic regression theory. This theory gives necessary and sufficient conditions, yields efficient algorithms for computing the NPMLE and leads us either directly to distribution theory or to rather specific conjectures about the asymptotic behavior.

Furthermore, the relation between NPMLE's and nonparametric least squares estimators will be discussed: these estimators actually coincide for interval censoring, case 1, but have a rather different behavior for interval censoring, case 2.

2 Interval censoring, case 1

We first discuss the following case of interval censoring.

Case 1. Let $(X_1, T_1), \ldots, (X_n, T_n)$ be a sample of random variables in $\mathbb{R}_+^2$, where X_i and T_i are independent (non-negative) random variables with distribution functions F_0 and G, respectively. The only observations which are available are T_i ("observation time") and $\delta_i = \{X_i \le T_i\}$. Here we denote the indicator of an event A (such as $\{X_i \le T_i\}$) just by A, instead of 1_A. The log likelihood for F_0 is given by the function

$$F \mapsto \sum_{i=1}^{n} \{\delta_i \log F(T_i) + (1-\delta_i)\log(1-F(T_i))\}, \tag{1}$$

where F is a right-continuous distribution function.

The (conditional) log likelihood, divided by n, can be written in the following way:

$$\psi(F) \stackrel{\text{def}}{=} \int_{\mathbb{R}^2} \left\{1_{\{x\le t\}} \log F(t) + 1_{\{x>t\}} \log\{1-F(t)\}\right\} dP_n(x,t), \tag{2}$$

where P_n is the empirical probability measure of the pairs (X_i, T_i), $1 \le i \le n$. The *nonparametric maximum likelihood estimator* (NPMLE) $\hat{F}_n$ of F is a (right-continuous) distribution function F, maximizing (2).

Remark 2.1. Note that only the values of $\hat{F}_n$ at the observation points matter for the maximization problem. To avoid trivialities, we will take as "the" NPMLE a distribution function which is piecewise constant, and only has jumps at the observation points. It may happen that the likelihood function is maximized by a function F such that $F(t) < 1$, at each observation point t. In this case we do not specify the location of the remaining mass to the right of the biggest observation point. Under these conventions, the NPMLE is uniquely determined, both in case 1 and case 2 of the interval censoring problem.

It turns out that in case 1 the NPMLE $\hat{F}_n$ coincides with the least squares estimator, obtained by minimizing the function

$$F \mapsto \sum_{i=1}^{n} (F(T_i) - \delta_i)^2$$

over the set of all distribution functions F (Remark 2.1 ensures uniqueness over the restricted class of dfs, having jumps only at the observation points).

Therefore the NPMLE is a straightforward solution of an isotonic regression problem; a fact that has already been used in the paper by AYER ET AL. (1955).

The pointwise asymptotic behavior of the NPMLE is studied in GROENEBOOM (1987) and the result is given again in GROENEBOOM AND WELLNER (1992) as Theorem 5.1:

Theorem 5.1 in Groeneboom and Wellner (1992). Let t_0 be such that $0 < F_0(t_0) < 1$, $0 < G(t_0) < 1$, and let F_0 and G be differentiable at t_0, with strictly positive derivatives $f_0(t_0)$ and $g(t_0)$, respectively. Furthermore, let $\hat{F}_n$ be the NPMLE of F_0. Then we have, as $n \to \infty$,

$$n^{1/3}\{\hat{F}_n(t_0) - F_0(t_0)\}/\{\tfrac{1}{2}F_0(t_0)(1 - F_0(t_0))f_0(t_0)/g(t_0)\}^{1/3} \xrightarrow{\mathcal{D}} 2Z,$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, and where Z is the last time where standard two-sided Brownian motion minus the parabola $y(t) = t^2$ reaches its maximum.

This shows that, under the conditions of the theorem, the NPMLE converges locally at the $n^{1/3}$ rate. A minimax result showing that the $n^{1/3}$ rate is the correct rate here and that the part of the constant in the minimax lower bound, depending on the underlying distribution, is correctly represented in the asymptotic variance of the NPMLE, is also shown in GROENEBOOM (1987) (in fact, two approaches are given there; one based on Assouad's Lemma and one based on the theory of limiting experiments, leading to slightly different universal constants in the lower bounds for the minimax risk). Still another proof of the minimax lower bound is sketched in the exercises of Chapter 2 of Part I of GROENEBOOM AND WELLNER (1992).

The minimax result was also recently reconsidered by GILL AND LEVIT (1992). Their approach is based on the van Trees inequality (VAN TREES (1968)). They recover the $n^{1/3}$ rate, but obtain a different type of constant, due to the fact that they use a (local) uniform Lipschitz condition on the underlying df (in contrast to the approach in GROENEBOOM (1987) and GROENEBOOM AND WELLNER (1992)).

As can be expected from the general theory on differentiable functionals (see e.g., VAN DER VAART (1991), efficient estimators of smooth functionals like the mean

$$\mu_{F_0} = \int t\, dF_0(t)$$

should have $\sqrt{n}$–behavior. Suppose that the support of P_{F_0} is a bounded interval $I = [0, M]$, and that F_0 and G have densities f_0 and g, respectively, satisfying

$$g(t) \geq \delta > 0, \text{ and } f_0(t) \geq \delta > 0, \qquad \text{if } t \in I,$$

for some $\delta > 0$. Further assume that g has a bounded derivative on I. An example of this situation is the case where F_0 and G are both the uniform distribution function on $[0, 1]$. Then we have the following result, proved in GROENEBOOM AND WELLNER (1992), Chapter 5 of Part II.

Theorem 5.5 in Groeneboom and Wellner (1992). Let F_0 and G satisfy the conditions, listed above, and let $\hat{F}_n$ be the NPMLE of F_0. Then

$$\sqrt{n}\int_I (\hat{F}_n(t) - F_0(t))\,dt \xrightarrow{\mathcal{D}} U,\ n \to \infty,$$

where U has a normal distribution with mean zero and variance

$$\sigma^2_{F_0} = \int \frac{F_0(t)(1 - F_0(t))}{g(t)}\,dt.$$

The proof uses a rather involved exponential martingale argument in order to give an upper bound to the probability that the maximum distance between successive jumps of $\hat{F}_n$ is bigger than $n^{-1/3}\log n$. This in turn is used to show that the supremum distance between $\hat{F}_n$ and F_0 is of order $n^{-1/3}\log n$. A different shorter proof, avoiding the upper bound argument for the supremum distance between $\hat{F}_n$ and F_0 and also treating more general functionals than the mean, is given in HUANG AND WELLNER (1995A).

The asymptotic variance of the above estimator of the mean is in fact the efficient asymptotic variance (i.e., coincides with the information lower bound) in this situation. Interestingly enough, the information lower bound calculation (done by Jon Wellner) preceded the result on the asymptotic variance of the estimator of the mean, based on the NPMLE. The lower bound calculation is given in VAN DER VAART (1991).

In the example on Hepatitis A in Bulgaria, given in KEIDING (1991), a quantity of interest is the *transmission potential* (i.e., the expected number of people infected by a person having the disease), which can be considered to be a smooth functional for a restricted class of distribution functions. In the model, used by KEIDING (1991), this quantity should be estimable at rate $n^{1/2}$ under smoothness conditions on the underlying distributions. Preliminary results on this are reported in HANSEN (1991). An intriguing aspect of the estimation of these global types of functionals is that the optimal bandwidth choice is quite different from the optimal bandwidth choice for the pointwise estimates.

3 Interval censoring, case 2

3.1 Characterization of the estimators

We now turn to the second case of interval censoring, mentioned in the introduction. From a mathematical (and possibly also practical) point of view this case is much more interesting than interval censoring, case 1. Much less is known, however, and the theory is still in its beginning stage. We consider the following model.

Interval censoring, Case 2. Let $(X_1,T_1,U_1),\ldots,(X_n,T_n,U_n)$ be a sample of random variables in $\mathbb{R}_+^3$, where X_i is a (non-negative) random variable with continuous distribution function F_0, and where T_i and U_i are (non-negative) random variables, independent of X_i, with a joint continuous distribution function H and such that $T_i \le U_i$ with probability one. The only observations which are available are (T_i,U_i) (the "observation times") and $\delta_i = \{X_i \le T_i\}$, $\gamma_i = \{X_i \in (T_i,U_i]\}$.

For a change, we start with discussing least squares estimators. A least squares estimator $\hat F_n$ of F_0 is defined as a minimizer of the function

$$F \mapsto \sum_{i=1}^n \Big\{w_{i,1}(F(T_i)-\delta_i)^2+w_{i,2}(F(U_i)-F(T_i)-\gamma_i)^2+w_{i,3}(F(U_i)-\gamma_i-\delta_i)^2\Big\}, \tag{3}$$

where the weights $w_{i,j}$ can be chosen in several different ways, to be discussed below. In different notation, we have to minimize

$$\psi(F) \stackrel{\mathrm{def}}{=} \int_{\mathbb{R}^3} \phi_F(x,t,u)\,dP_n(x,t,u), \tag{4}$$

where

$$\begin{aligned}\phi_F(x,t,u) \stackrel{\mathrm{def}}{=}\; & w_1(t,u)(F(t)-1_{\{x\le t\}})^2\\ & +w_2(t,u)(F(u)-F(t)-1_{\{t<x\le u\}})^2\\ & +w_3(t,u)(1-F(u)-1_{\{x>u\}})^2,\end{aligned} \tag{5}$$

and P_n is the empirical probability measure of the triples (X_i,T_i,U_i), $1 \le i \le n$; the weight functions w_j, $j=1,2,3$, only have to be defined at the points (T_i,U_i) by

$$w_j(T_i,U_i) = w_{i,j},\; i=1,\ldots,n;\; j=1,2,3.$$

where $w_{i,j}$ is defined as in (5).

Remark 3.1. Note that again (as in the preceding section) only the values of $\hat F_n$ at the observation points T_i and U_i matter for the minimization problem. We will take as "the" least squares estimator a distribution function which is piecewise constant, and only has jumps at the observation points T_i and U_i. It may again happen that the function ψ is minimized by a function F such that $F(t) < 1$, at each observation point t. In this case we do not specify the location of the remaining mass to the right of the biggest observation point. We shall show that, under these conventions, the least squares estimator is uniquely determined.

We start by characterizing the least squares estimator, under the conventions of Remark 3.1. To this end, we introduce the following processes.

Definition 3.1 Let F be a distribution function on $[0,\infty)$. Then the process W_F is defined by

$$\begin{aligned} W_F(t) = \quad & \int_{t'\in[0,t]} w_1(t',u)\{1_{\{x\le t'\}} - F(t')\}\, dP_n(x,t',u) \\ & - \int_{t'\in[0,t]} w_2(t',u)\Big\{1_{\{t'<x\le u\}} - (F(u)-F(t'))\Big\}\, dP_n(x,t',u) \\ & + \int_{u\in[0,t]} w_2(t',u)\Big\{1_{\{t'<x\le u\}} - (F(u)-F(t'))\Big\}\, dP_n(x,t',u) \\ & - \int_{u\in[0,t]} w_3(t',u)\Big\{1_{\{x>u\}} - (1-F(u))\Big\}\, dP_n(x,t',u), \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{for } t\ge 0, \end{aligned} \tag{6}$$

where P_n is the empirical probability measure of the points (X_i,T_i,U_i), $i = 1,\ldots,n$.

The following proposition characterizes the least squares estimator.

Proposition 1 *Let $\mathcal{F}$ be the set of discrete distribution functions, with mass concentrated at the observation points and possibly some extra mass at the right of the biggest observation point. Then $\hat{F}_n$ minimizes the right-hand side of (3) over all $F\in\mathcal{F}$ if and only if*

$$\int_{[t,\infty)} dW_{\hat{F}_n}(t') \le 0, \qquad \forall t\ge 0, \tag{7}$$

and

$$\int \hat{F}_n(t)\, dW_{\hat{F}_n}(t) = 0, \tag{8}$$

where W_F is defined.by (6). Moreover, $\hat{F}_n$ is uniquely determined by (7) and (8).

The proof is quite similar to the proof of Proposition 1.3 in Chapter 1, part II, of Groeneboom and Wellner (1992), but slightly easier, since we don't have to worry about the endpoints, which caused some extra work in the characterization of the NPMLE. In order to describe an algorithm for computing the least squares estimator, we introduce a "time scale process" similar to (but different from) the time scale process G_F, defined by (1.29) in Chapter 1, part II, of Groeneboom and Wellner (1992).

Definition 3.2. Let F be a distribution function on $[0,\infty)$ and let H_n be the empirical distribution function of the pairs (T_i, U_i). Then the processes G and V_F are defined by

$$\begin{aligned} G(t) = \quad & \int_{t'\in[0,t]}\{w_1(t',u)+w_2(t',u)\}\,dH_n(t',u) \\ & + \int_{u\in[0,t]}\{w_2(t',u)+w_3(t',u)\}\,dH_n(t',u),\ t\ge 0, \end{aligned} \tag{9}$$

and

$$V_F(t) = W_F(t) + \int_{[0,t]} F(t')\,dG(t'),\ t \ge 0. \tag{10}$$

The processes G and V_F have similar motivation and properties as the processes G_F and V_F on page 49 of Groeneboom and Wellner (1992).

The following proposition characterizes $\hat F_n$ as the slope of the convex minorant of a self-induced cumulative sum diagram.

Proposition 2 *Let the class of distribution functions $\mathcal{F}$ be defined as in Proposition 1. Then $\hat F_n$ minimizes the right-hand side of (3) over $\mathcal{F}$ if and only if $\hat F_n$ is the left derivative of the convex minorant of the "cumulative sum (cusum) diagram", consisting of the points*

$$P_j = \left(G(T_{(j)}), V_{\hat F_n}(T_{(j)})\right),$$

where $P_0 = (0,0)$ and $T_{(j)}$, $j = 1,2,\ldots,2n$, are the ordered observation times.

This suggests a simple iterative procedure for computing the least squares estimator: starting with an arbitrary (sub)distribution function, one computes at the $(m+1)^{th}$ iteration step the convex minorant of the cusum diagram, consisting of the points

$$P_j = \left(G(T_{(j)}), V_{F^{(m)}}(T_{(j)})\right),$$

and uses the left derivative $F^{(m+1)}$ of the convex minorant in the process $V_{F^{(m+1)}}$, defining the cusum diagram in the next iteration. We will show in the next section that this procedure will converge to the solution from any starting distribution.

The NPMLE can in this case be characterized as a least squares estimator with "self-induced weights". In fact, the NPMLE is characterized by Proposition 1, but with the weights w_i in the process W_F in (6) defined by

$$w_1(t,u) = 1/F(t),\ w_2(t,u) = 1/(F(u)-F(t)),\ \text{and}\ w_3(t,u) = 1/(1-F(u)). \tag{11}$$

If a denominator in (11) equals zero, the corresponding weight is defined to be infinite and the corresponding squared distance in (5) is equal to zero in that case. Using the convention $0 \cdot \infty = 0$, the corresponding weighted square gives no contribution to the total sum of squares in (3). In practice, one actually performs a preliminary reduction of the problem, excluding these terms from the minimization problem.

So in this case the weights are defined by the solution itself, a situation somewhat reminiscent of the "self-consistency equations". In an iterative convex minorant algorithm, the weights are adjusted in an iterative procedure in such a way that the solution and the weights match at the end of the iteration.

3.2 Algorithms

We show that the iterative convex minorant algorithm, based on Proposition 2, corresponds to a contraction mapping for a suitably chosen norm on $\mathcal{F}$, with a contraction constant depending on the weight function. Since there is only one fixed point, the algorithm will converge from any starting point.

We define the L_2–distance $\|\cdot\|$ on $\mathcal{F}$ by

$$\|F_1 - F_2\|^2 = \int (F_1(t) - F_2(t))^2 \, dG(t),$$

where G is defined by (9). Let the function

$$t \mapsto \frac{dV_F}{dG}(t)$$

be defined by

$$\frac{dV_F}{dG}(t) = \begin{cases} \frac{V_F(t)-V_F(t-)}{G(t)-G(t-)} & , \text{ if } G(t) > G(t-), \\ 0 & , \text{ otherwise.} \end{cases}$$

We define $F^{(m+1)}$ at the $(m+1)^{th}$ iteration step as the distribution function in $\mathcal{F}$ that minimizes

$$F \mapsto \left\| F - \frac{dV_{F^{(m)}}}{dG} \right\|.$$

Let the mapping $T : F \mapsto T_F$, $F \in \mathcal{F}$ be defined by

$$\left\| T_F - \frac{dV_F}{dG} \right\| = \min_{F' \in \mathcal{F}} \left\| F' - \frac{dV_F}{dG} \right\|.$$

Then, by Theorem 8.2.5 in Robertson, Wright and Dykstra (1988),

$$\left\| F^{(m+1)} - F^{(m)} \right\| = \left\| T(F^{(m)}) - T(F^{(m-1)}) \right\| \le \left\| \frac{dV_{F^{(m)}}}{dG} - \frac{dV_{F^{(m-1)}}}{dG} \right\|. \quad (12)$$

But the square of the term at the right-hand side of (12) can be written

$$\begin{aligned}&\int\{F^{(m)}(u)-F^{(m-1)}(u)\}^2\,\tfrac{w_2(t',u)^2}{w_1(t',u)+w_2(t',u)}\,dH_n(t',u)\\&\qquad+\int\{F^{(m)}(t')-F^{(m-1)}(t')\}^2\tfrac{w_2(t',u)^2}{w_2(t',u)+w_3(t',u)}\,dH_n(t',u)\\&\le c\cdot\left\|F^{(m)}-F^{(m-1)}\right\|^2,\end{aligned}\tag{13}$$

where the constant c satisfies

$$c\le\max_{1\le i\le n}\max\left\{\frac{w_2(T_i,U_i)^2}{(w_1(T_i,U_i)+w_2(T_i,U_i))^2},\frac{w_2(T_i,U_i)^2}{(w_2(T_i,U_i)+w_3(T_i,U_i))^2}\right\}<1$$

As an example, if $w_i(t,u)\equiv 1$, $i=1,2,3$, we get

$$\left\|F^{(m+1)}-F^{(m)}\right\|\le\tfrac12\left\|F^{(m)}-F^{(m-1)}\right\|.$$

For finding the NPMLE one could carry out the iteration procedure above repeatedly, for example starting with equal weights. This amounts to a repeated weighted least squares procedure, where the weights are determined by the preceding step. At the start of each iteration after the initial iteration one takes the weights as in (5), but with F defined as the solution of the least squares problem in the preceding step. A program for doing this (using some "buffers", preventing the iterative estimates from leaving the allowed region) has been developed and seems to work fine. Another (simpler) iterative convex minorant algorithm for computing the NPMLE is discussed in GROENEBOOM AND WELLNER (1992), Chapter 3 of Part II. It is shown in JONGBLOED (1995A) and JONGBLOED (1995B) that a slight modification of the latter algorithm will always converge.

However, the original motivation for developing these algorithms was an attempt to derive distribution theory. We will turn to this in the next section.

3.3 Local distribution theory for case 2

For interval censoring, case 1, we have the result that the NPMLE converges at rate $n^{1/3}$. Interestingly enough, in case 2 there exist estimators which have a faster rate of convergence. First of all, a minimax calculation shows that the rate of convergence should not be $n^{1/3}$ but $(n\log n)^{1/3}$. The lower bound calculation is given in BAKKER (1988). GILL AND LEVIT (1992) also derive a lower bound of order $(n\log n)^{-1/3}$. A simple histogram-type estimator has been constructed by Lucien Birgé (personal communication), which can easily be shown to attain the rate $(n\log n)^{1/3}$ at t_0. The trouble with the least squares estimator with constant weights is that observations lying in smaller intervals do not get more weight; they should get more weight in order to obtain the faster rate of convergence!

It is conjectured that the least squares estimator with weights, inversely proportional to the lengths of the observation intervals, converges locally at rate $(n \log n)^{1/3}$. Computer experiments also point in this direction. What in our view is actually more interesting is that the NPMLE seems to behave asymptotically as a least squares estimator with weights w_i defined by

$$w_i(t,u) = 1/(F_0(u) - F_0(t)).$$

In fact there exist now a group of connected conjectures about the behavior of the NPMLE, all pointing in the direction of the following conjecture.

Conjecture. Let F_0 and H be continuously differentiable at t_0 and (t_0, t_0), respectively, with strictly positive derivatives $f_0(t_0)$ and $h(t_0, t_0)$. By continuous differentiability of H at (t_0, t_0) is meant that the density $h(t, u)$ is continuous in (t, u) if $t < u$ and (t, u) is sufficiently close to (t_0, t_0) and that $h(t, t)$, defined by

$$h(t,t) = \lim_{u \downarrow t} h(t,u),$$

is continuous in t, for t in a neighborhood of t_0.

Let $0 < F_0(t_0), H(t_0, t_0) < 1$, and let $\hat{F}_n$ be the NPMLE. Then

$$(n \log n)^{1/3} \{\hat{F}_n(t_0) - F_0(t_0)\} / \{\tfrac{3}{4} f_0(t_0)^2 / h(t_0, t_0)\}^{1/3} \xrightarrow{\mathcal{D}} 2Z,$$

where Z is the last time where standard two-sided Brownian motion minus the parabola $y(t) = t^2$ reaches its maximum.

The conjecture is discussed in Part II, Chapter 5, section 2, of Groeneboom and Wellner (1992), where a result of this type is proved for an estimator, obtained after one step of an iterative convex minorant algorithm, starting with the underlying distribution. Of course, for practical purposes the latter result is useless; the study of its behavior was only motivated by the belief that its behavior is the clue to the behavior of the NPMLE.

4 Estimation of smooth functionals

4.1 Information lower bounds

As was remarked earlier, one can expect that smooth functionals of the model can be estimated at $\sqrt{n}$-rate. The theory on the estimation of smooth functionals for case 2 is rather complicated, though, and intimately connected with certain Fredholm integral equations for which solutions can only be given implicitly. We will give a sketch of the present situation of the theory below, relying mostly on the exposition in GESKUS AND GROENEBOOM (1995A,B,C).

For a more complete and more general treatise on the relation between pathwise differentiability of functionals and asymptotic efficiency, we refer to part I of (Groeneboom and Wellner (1992)) or (Bickel *et. al.* (1993)). We give some key concepts below.

Let the unknown distribution P on the space $(\mathcal{Y}, \mathcal{B})$ be contained in some class of probability measures $\mathcal{P}$, which is dominated by a σ-finite measure μ. Let P have density p with respect to μ. Since we are interested in estimation of some real-valued function of P, we introduce the functional $\Theta : \mathcal{P} \to \mathbb{R}$. Let, for some $\delta > 0$, the collection $\{P_t\}$ with $t \in (0, \delta)$ be a one-dimensional parametric submodel which is smooth in the following sense:

$$\int \left[t^{-1}(\sqrt{p_t} - \sqrt{p}) - \tfrac{1}{2}a\sqrt{p}\right]^2 d\mu \to 0 \quad \text{as } t \downarrow 0, \text{ for some } a \in L_2(P)$$

Such a submodel is called *Hellinger differentiable* and a is called the *score function* or *score*. The folowing result is well-known.

Proposition 3 *Each score belonging to some Hellinger differentiable submodel is contained in* $L_2^0(P)$.

Proof: See GESKUS AND GROENEBOOM (1995C) □

In our situation the collection of scores a, obtained by considering all possible one-dimensional Hellinger-differentiable parametric submodels, is a linear space. This space is called the *tangent space* at P, denoted by $T(P)$. Note that $T(P) \subset L_2^0(P)$.

Now $\Theta : \mathcal{P} \to \mathbb{R}$ is pathwise differentiable at P if for each Hellinger differentiable path $\{P_t\}$, with corresponding score a, we have

$$\lim_{t \downarrow 0} t^{-1}(\Theta(P_t) - \Theta(P)) = \Theta'_P(a),$$

$$\text{with } \Theta'_P : T(P) \to \mathbb{R} \text{ continuous and linear.}$$

Θ'_P can be written in an inner product form. Since $T(P)$ is a subspace of the Hilbert-space $L_2(P)$, the continuous linear functional Θ'_P can be extended to a continuous linear functional $\overline{\Theta}'_P$ on $L_2(P)$. By the Riesz representation theorem, to $\overline{\Theta}'_P$ belongs a unique $\theta_P \in L_2(P)$, called the *gradient*, satisfying

$$\overline{\Theta}'_P(h) = <\theta_P, h>_P \quad \text{for all } h \in L_2(P).$$

One gradient is playing a special role, which is obtained by extending $T(P)$ to the Hilbert space $\overline{T(P)}$. Then, the extension of Θ'_P *is* unique, yielding the *canonical gradient* or *efficient influence function* $\tilde{\theta}_P \in \overline{T(P)}$. This canonical gradient is also obtained by taking the orthogonal projection of any gradient θ_P, obtained after extension of Θ'_P, into $\overline{T(P)}$. Hence $\tilde{\theta}_P$ is the gradient with minimal norm among all gradients and we have

$$\|\theta_P\|_P^2 = \|\tilde{\theta}_P\|_P^2 + \|\theta_P - \tilde{\theta}_P\|_P^2.$$

The so-called convolution theorem now says that the smallest asymptotic variance we can get for a regular estimator of $\Theta(P)$ is $\|\tilde{\theta}_P\|^2$. An *asymptotically efficient* estimator is a regular estimator which has an asymptotic distribution with this (minimal) variance.

The interval censoring model is an example of a model with information loss, in which the distribution P is induced by a transformation. In these models the functional to be estimated is implicitly defined. The lower bound theory for such implicitly defined functionals is treated in VAN DER VAART (1991) and BICKEL *et. al.* (1993). This theory will be applied to case 2 of the interval censoring model. We start with the formulation of the model for case 2. The loss of information is expressed by the fact that, instead of a sample $(X_1, \ldots, X_n)$, we observe $(T_1, U_1, \Delta_1, \Gamma_1), \ldots, (T_n, U_n, \Delta_n, \Gamma_n)$ with $\Delta_i = 1_{\{X_i \le T_i\}}$ and $\Gamma_i = 1_{\{T_i < X_i \le U_i\}}$. We suppose:

(M1) X_i is a non-negative absolutely continuous random variable with distribution function F. Let $S > 0$. F is contained in the class

$$\mathcal{F}_S := \{F \mid \text{support}(F) \subset [0, S];\ F \ll \lambda,\ \lambda \text{ being Lebesgue measure}\}.$$

F is the distribution on which we want to obtain information; however, we do not observe X_i directly.

(M2) Instead, we observe the pairs (T_i, U_i), with distribution function H. H is contained in $\mathcal{H}$, the collection of all two-dimensional distributions on $\{(t,u) \mid 0 \le t < u\}$, absolutely continuous with respect to two-dimensional Lebesgue measure and such that each H is independent of each F. Let h denote the density of (T_i, U_i), with marginal densities and distribution functions h_1, H_1 and h_2, H_2 for T_i and U_i respectively.

(M3) If both H_1 and H_2 put zero mass on some set A, then F has zero mass on A as well, so $F \ll H_1 + H_2$. This means that F does not have mass on sets in which no observations can occur.

Condition (M3) is needed to ensure consistency. Moreover, without this assumption the functionals we are interested in are not well-defined. So discrete F should be excluded from $\mathcal{F}_S$.

Note that what we *do* observe can be seen as a measurable transformation S of what we *would* observe if there would be no censoring:

$$S(x,t,u) = (t,u,\delta,\gamma),$$

with domain $\{(x,t,u)\,|\,0 \le x,\ 0 \le t < u\}$. This domain will be called the hidden space, and the image space will be called the observation space. In our model P is induced by F and H, and is from now on written as $Q_{F,H}$, having density

$$q_{F,H}(t,u,\delta,\gamma) = h(t,u)F(t)^{\delta}(F(u)-F(t))^{\gamma}(1-F(u))^{1-\delta-\gamma}$$

with respect to $\lambda_2 \otimes \nu_2$, where ν_2 denotes the counting measure on the set $\{(0,1),(1,0),(0,0)\}$.

We are interested in estimation of some functional $K(F)$ of F. However, $K(F)$ is only implicitly defined as $\Theta(Q_{F,H})$, with H acting as a nuisance parameter. In particular, we will be concerned with the problem whether the NPMLE $\hat{\Theta}_n$ of $\Theta(Q_{F,H})$ satisfies

$$\sqrt{n}(\hat{\Theta}_n - \Theta(Q_{F,H})) \xrightarrow{\mathcal{D}} N(0, \|\tilde{\theta}_{Q_{F,H}}\|^2).$$

All Hellinger differentiable submodels at $Q_{F,H}$ that can be formed, together with the corresponding score functions, are induced by the Hellinger differentiable paths of densities on the hidden space, according to the following theorem:

Theorem 4.1 *Let $\mathcal{P} \ll \mu$ be a class of probability measures on the hidden space $(\mathcal{Y},\mathcal{B})$. $P \in \mathcal{P}$ is induced by the random vector Y. Suppose that the path $\{P_t\}$ to P satisfies*

$$\int \left[t^{-1}(\sqrt{p_t}-\sqrt{p}) - \tfrac{1}{2}a\sqrt{p}\right]^2 d\mu \to 0 \quad \text{as } t \downarrow 0$$

for some $a \in L_2^0(P)$.
Let $S : (\mathcal{Y},\mathcal{B}) \to (\mathcal{Z},\mathcal{C})$ be a measurable mapping. Suppose that the induced measures $Q_t = P_tS^{-1}$ and $Q = PS^{-1}$ on $(\mathcal{Z},\mathcal{C})$ are absolutely continuous with respect to μS^{-1}, with densities q_t and q. Then the path $\{Q_t\}$ is also Hellinger differentiable, satisfying

$$\int \left[t^{-1}(\sqrt{q_t}-\sqrt{q}) - \tfrac{1}{2}\overline{a}\sqrt{q}\right]^2 d\mu S^{-1} \to 0 \quad \text{as } t \downarrow 0$$

with $\overline{a}(z) = E_P(a(Y)|S=z)$.

Proof: See Bickel *et. al.* (1993).

Note that $\overline{a} \in L_2^0(Q)$. The relation between the scores a in the hidden tangent space $T(P)$ and the induced scores $\overline{a}$ is expressed by the mapping

$$A_P : a(\cdot) \mapsto E_P(a(Y)|S = \cdot).$$

This mapping is called the *score operator.* It is continuous and linear. Its range is the induced tangent space, which is contained in $L_2^0(Q)$.

Now Theorem 4.1 yields the tangent space $T(Q_{F,H})$ of the induced Hellinger differentiable paths $\{Q_t\}$ at $Q_{F,H}$ with score operator $A : L_2^0(F) \oplus L_2^0(H) \to T(Q_{F,H})$ given by:

$$[A_{F,H}(a+e)](t,u,\delta,\gamma) = E_{F,H}\{a(X) + e(T,U) \,|\, (T,U,\Delta,\Gamma) = (t,u,\delta,\gamma)\}$$

Having specified the Hellinger differentiable paths in the observation space, we can also determine differentiability of the functional

$$\Theta(Q_{F,H}) = K(F).$$

Note that $\Theta(Q_{F,H})$ is defined unambiguously by condition (M3).

In our censoring model, differentiability of $\Theta(Q_{F,H})$ along the induced Hellinger differentiable paths in the observation space can be proved by looking at the structure of the adjoint $A^*_{F,H}$ of the map $A_{F,H}$ according to Theorem 4.2 below, which was first proved in VAN DER VAART (1991) in a more general setting, allowing for Banach space valued functions as estimand. Then the proof is slightly more elaborate.

Recall that the adjoint of a continuous linear mapping $A : D \to E$, with D and E Hilbert-spaces, is the unique continuous linear mapping $A^* : D \to E$ satisfying

$$< Ag, h >_E = < g, A^*h >_D \quad \forall g \in G, h \in H.$$

The score operator from Theorem 4.1 is playing the role of A. Its adjoint can be written as a conditional expectation as well. If $Z \sim PS^{-1}$, then:

$$[A_P^* b](y) = E_P(b(Z)|Y = y) \quad \text{a.e.-}[P]$$

Theorem 4.2 *Let $\mathcal{Q} = \mathcal{P}S^{-1}$ be a class of probability measures on the image space of the measurable transformation S. Suppose the functional $\Theta : \mathcal{Q} \to \mathbb{R}$ can be written as $\Theta(Q_P) = K(P)$ with K pathwise differentiable at P in the hidden space, having canonical gradient $\tilde{\kappa}_P$.*
Then Θ is differentiable at $Q_P \in \mathcal{Q}$ along the collection of induced paths in the observation space obtained via Theorem 4.1 if and only if

$$\tilde{\kappa}_P \in \mathcal{R}(A_P^*) \tag{14}$$

If (14) holds, then the canonical gradients $\tilde{\theta}_{Q_P}$ of Θ and $\tilde{\kappa}_P$ of K are related by

$$\tilde{\kappa}_P = A_P^* \tilde{\theta}_{Q_P}$$

Proof: See VAN DER VAART (1991) or GESKUS AND GROENEBOOM (1995C).

□

Now $K(F)$ is only implicitly defined as $\Theta(Q_{F,H})$, with H acting as a nuisance parameter. Note that $\Theta(Q_{F,H})$ is defined unambiguously by condition (M3). The key equation that is needed is the following

$$\tilde{\kappa}_F \in \mathcal{R}(L_1^*)$$

and if this holds, then the canonical gradient is the unique element $\tilde{\theta}$ in $\overline{\mathcal{R}(L_1)}$ satisfying

$$L_1^* \tilde{\theta} = \tilde{\kappa}_F. \tag{15}$$

The operators L_1 and L_2 have the following form:

$$\begin{aligned} &[L_1 a](u,v,\delta,\gamma) = \tfrac{\delta \int_0^u a\,dF}{F(u)} + \tfrac{\gamma \int_u^v a\,dF}{F(v)-F(u)} + \tfrac{(1-\delta-\gamma)\int_v^M a\,dF}{1-F(v)} \quad &\text{a.e.} - [Q_{F,H}] \\ &[L_2 e](u,v,\delta,\gamma) = e(u,v) &\text{a.e.} - [Q_{F,H}] \end{aligned} \tag{16}$$

The adjoint of L_1 can be written as $[L_1^* b](x) = E_P(b(U,V,\Delta,\Gamma)|X = x)$ and we get

$$\begin{aligned} [L_1^* b](x) \;=\; & \int_{t=x}^M \int_{u=t}^M b(t,u,1,0)\,h(t,u)\,dt\,du + \\ & \int_{t=0}^x \int_{u=x}^M b(t,u,0,1)\,h(t,u)\,dt\,du + \\ & \int_{t=0}^x \int_{u=t}^x b(t,u,0,0)\,h(t,u)\,dt\,du \qquad \text{a.e.-}[F]. \end{aligned} \tag{17}$$

Many functionals that are pathwise differentiable in the model without censoring, lose this property in the interval censoring model. Any functional K with a canonical gradient that is not a.e. equal to a continuous function cannot be obtained under L_1^*. So not all linear functionals remain pathwise differentiable. For example, $\kappa(F) = F(t_0)$, with canonical gradient $1_{[0,t_0]}(\cdot) - F(t_0)$, loses this property. This is in correspondence with $F(t_0)$ not being estimable at $\sqrt{n}$-rate. However, functionals of the form $K(F) = \int c(x)\,dF(x)$, with c sufficiently smooth, can be shown to remain differentiable under censoring. Hence for these functionals the above information lower bound theory holds.

We will be concerned with the problem whether the NPMLE $\hat{\Theta}_n$ of $\Theta(Q_{F,H})$ satisfies

$$\sqrt{n}(\hat{\Theta}_n - \Theta(Q_{F,H})) \xrightarrow{\mathcal{D}} N(0, \|\tilde{\theta}_{Q_{F,H}}\|^2).$$

In the interval censoring model, both case 1 and case 2, the function

$$\phi(x) := \int_x^M a(t)\,dF(t) \text{ with } a \in L_2^0(F).$$

appears explicitly in the score operator L_1. Therefore it plays an important role. It is called the *integrated score function.* ¿From its definition we know that ϕ satisfies $\phi(0)=\phi(M)=0$ and that ϕ is continuous for $F \in \mathcal{F}_S$.

We now investigate solvability of the equation

$$\tilde{\kappa}_F = L_1^* L_1 a$$

in the variable $a \in L_2^0(F)$. By the structure of the score operator L_1 this can be reformulated as an equation in ϕ:

$$\begin{aligned} \tilde{\kappa}_F(x) \quad = \quad & \int_{t=0}^{x}\int_{u=t}^{x} \tfrac{\phi(u)}{1-F(u)}\, h(t,u)\,du\,dt \\ & - \int_{t=0}^{x}\int_{u=x}^{M} \tfrac{\phi(u)-\phi(t)}{F(u)-F(t)}\, h(t,u)\,du\,dt \\ & - \int_{t=x}^{M}\int_{u=t}^{M} \tfrac{\phi(t)}{F(t)}\, h(t,u)\,du\,dt \qquad \text{a.e.-}[F]. \end{aligned} \tag{18}$$

The support of F may consist of several disjoint intervals. However, (18) is not defined on intervals where F does not put mass, and these intervals do not play any role. So without loss of generality we may assume the support of F to consist of one interval $[0, M]$.

Unlike case 1, differentiating equation (18) on both sides does not yield an explicit formula for ϕ. Instead, we get the following integral equation:

$$\phi(x)+d_F(x)\left[\int_{t=0}^{x} \tfrac{\phi(x)-\phi(t)}{F(x)-F(t)}\, h(t,x)\,dt - \int_{t=x}^{M} \tfrac{\phi(t)-\phi(x)}{F(t)-F(x)}\, h(x,t)\,dt\right] = k(x)d_F(x), \tag{19}$$

with $d_F(x)$ being the function

$$d_F(x) = \tfrac{F(x)[1-F(x)]}{h_1(x)[1-F(x)]+h_2(x)F(x)},$$

writing $k(x)$ instead of $\tilde{\kappa}'_F(x)$. Although k may depend on the underlying distribution, we do not explicitly express this dependence. Apart from the model conditions $(M1)$ to $(M3)$, some extra conditions will have to be introduced.

(S1) h_1 and h_2 are continuous, with $h_1(x) + h_2(x) > 0$ for all $x \in [0, M]$.

(S2) $h(t, u)$ is continuous

(S3) $\text{Prob}\{U - T < \epsilon_0\} = 0$ for some ϵ_0 with $0 < \epsilon_0 \le 1/2\, M$, so h does not have mass close to the diagonal

(S4) F is either a continuous distribution function with support $[0, M]$, or a piecewise constant distribution function with a finite number of jumps, all in $[0, M]$; F satisfies

$$F(u) - F(t) \ge c > 0, \text{ if } u - t \ge \epsilon_0$$

(S5) k is continuous

The integral equation for ϕ belongs to a well-known family of integral equations, which have been studied extensively, the family of *Fredholm integral equations of the second kind.* Using this theory, it is proved that equations (19) have a (unique) solution. If we impose some extra smoothness conditions, we can derive some smoothness properties of the solution. These smoothness properties also imply solvability of $\tilde{\kappa}_F = L_1^* L_1 a$ for the unknown absolutely continuous distribution function F. The extra smoothness conditions are:

(L1) The partial derivatives $\Delta_x^1(t) = \frac{\partial}{\partial x} h(t, x)$ and $\Delta_x^2(t) = \frac{\partial}{\partial x} h(x, t)$ exist, except for at most a countable number of points x, where left and right derivatives exist. The derivatives are bounded, uniformly over t and x.

(L2) k is differentiable, except for at most a countable number of points x, where left and right derivatives exist. The derivative is bounded, uniformly over x.

We now can specify the structure of the canonical gradient $\tilde{\theta}_F \in \mathcal{R}(L_1)$:

$$\tilde{\theta}_F(t, u, \delta, \gamma) = -\delta\, \frac{\phi_F(t)}{F(t)} - \gamma\, \frac{\phi_F(u) - \phi_F(t)}{F(u) - F(t)} + (1 - \delta - \gamma)\, \frac{\phi_F(u)}{1 - F(u)}, \quad (20)$$

where ϕ_F satisfies the integral equation (19).

4.2 Asymptotic efficiency of the NPMLE

In this section, we will denote the underlying distribution function by F_0. Under uniqueness, proposition 1.3 in GROENEBOOM AND WELLNER (1992) gives an alternative criterion which is necessary and sufficient for the NPMLE.

Given a sample $(U_1, V_1, \Delta_1, \Gamma_1), \ldots, (U_n, V_n, \Delta_n, \Gamma_n)$, let $\mathcal{F}$ be the class of distribution functions F satisfying

$$\begin{cases} F(U_i) > 0 & , \text{ if } X_i \le U_i, \\ F(V_i) - F(U_i) > 0 & , \text{ if } U_i < X_i \le V_i, \\ 1 - F(V_i) > 0 & , \text{ if } X_i > V_i, \end{cases}$$

and having mass concentrated on the set of observation points augmented with an extra point bigger than all observation points. It is easily seen that $\hat{F}_n$ belongs to this class. For distribution functions $F \in \mathcal{F}$, the following process $t \mapsto W_F(t)$ is properly defined:

$$\begin{aligned} W_F(t) = \quad & \int_{u\in[0,t]} \delta\, F(u)^{-1} dQ_n(u,v,\delta,\gamma) \\ & - \int_{u\in[0,t]} \gamma\, \{F(v) - F(u)\}^{-1} dQ_n(u,v,\delta,\gamma) \\ & + \int_{v\in[0,t]} \gamma\, \{F(v) - F(u)\}^{-1} dQ_n(u,v,\delta,\gamma) \\ & - \int_{v\in[0,t]} (1-\delta-\gamma)\, \{1 - F(v)\}^{-1} dQ_n(u,v,\delta,\gamma), \\ & \qquad\qquad\qquad\qquad \text{for } t \ge 0, \end{aligned} \tag{21}$$

where Q_n is the empirical probability measure of the points $(U_i, V_i, \Delta_i, \Gamma_i)$, $i = 1, \ldots, n$.

Let $J_i = [\tau_{i-1}, \tau_i)$, $i = 1, \ldots, k+1$, $\tau_0 = 0$, $\tau_{k+1} = M$ and τ_i is a point of jump of $\hat{F}_n$, $i = 1, \ldots, k$. So τ_1 and τ_k are the first and last point of jump of $\hat{F}_n$ respectively. Restriction to a compact interval $[0, M]$ is only needed to obtain the efficiency result Theorem 4.3, but not needed for Proposition 4, Corollary 4.1 and the consistency result (24).

Now proposition 1.3 in Groeneboom and Wellner (1992) says

Proposition 4 *The function $\hat{F}_n$ maximizes the likelihood over all $F \in \mathcal{F}$ if and only if*

$$\int_{[t,\tau_k]} dW_{\hat{F}_n}(t') \le 0, \qquad \forall t \ge \tau_1, \tag{22}$$

and

$$\int_{[\tau_1,\tau_k]} \hat{F}_n(t)\, dW_{\hat{F}_n}(t) = 0. \tag{23}$$

Moreover, $\hat{F}_n$ is uniquely determined by (22) and (23).

Note that there may be observation points before τ_1 and beyond τ_k. However, there the NPMLE should be 0 and 1 respectively. (See the discussion before proposition 1.3 in Groeneboom and Wellner (1992).) Now the following corollary, proved in Geskus and Groeneboom (1995b) is an immediate consequence.

Corollary 4.1 *Any function σ that is constant on the same intervals as $\hat{F}_n$ satisfies*

$$\int_{u\in J_i} \sigma(u)\left\{\frac{\delta}{\hat{F}_n(u)} - \frac{\gamma}{\hat{F}_n(v)-\hat{F}_n(u)}\right\} dQ_n(u,v,\delta,\gamma)$$
$$+\int_{v\in J_i} \sigma(v)\left\{\frac{\gamma}{\hat{F}_n(v)-\hat{F}_n(u)} - \frac{1-\gamma-\delta}{1-\hat{F}_n(v)}\right\} dQ_n(u,v,\delta,\gamma) = 0$$

for $i = 2,\ldots,k$.

Remark. In fact corollary 4.1 follows from Fenchel duality theory (see e.g. ROCKAFELLAR (1970), theorem 28.3).

Moreover we have uniform consistency of the NPMLE of F_0 (see GROENEBOOM AND WELLNER (1992), part II, section 4.3):

$$\mathrm{Prob}\left\{\lim_{n\to\infty} \|\hat{F}_n - F_0\|_\infty = 0\right\} = 1 \tag{24}$$

Another result that will be needed can be deduced from VAN DE GEER (1993).

Lemma 4.1 *For $i = 1, 2$,*

$$\|\hat{F}_n - F_0\|_{H_i} = \mathcal{O}_p(n^{-1/3}(\log n)^{1/6}) \text{ as } n \to \infty,$$

where H_1 and H_2 are the first and second marginal distribution function of H, respectively.

In order to be able to use Lemma 4.1 one further specification is made to the kind of functionals that are allowed:

(D1) $$K(G) - K(F_0) = \int \tilde{\kappa}(x)\, d(G-)(x) + \mathcal{O}(\|G - F_0\|_2^2),$$

for all distribution functions G with support contained in $[0, M]$, and where $\|G - F_0\|_2$ is the L_2-distance between the distribution functions G and F_0 w.r.t. Lebesgue measure on $\mathbb{R}$.

We also make the following assumption:

(D2) The underlying distribution function F_0 has a density bounded away from zero.

By condition (D2) and the strong consistency of the NPMLE, there exists a constant c, such that

$$\hat{F}_n(u) - \hat{F}_n(t) \ge c, \text{ if } u - t \ge \epsilon_0, \tag{25}$$

if n is sufficiently large.

Combining all preceding results we then obtain the following theorem (Theorem 2.1 in GESKUS AND GROENEBOOM (1995B)), showing efficiency of the NPMLE:

Theorem 4.3 *Let the following conditions on F_0, H and $\tilde{\kappa}_{F_0}$ be satisfied: (M1) to (M3),(S1) to (S5), (L1) and (L2) of the preceding section, and (D1) and (D2).*
Then we have

$$\sqrt{n}(K(\hat{F}_n) - K(F_0)) \xrightarrow{\mathcal{D}} N(0, \|\tilde{\theta}\|^2_{Q_{F_0}}) \quad as \quad n \to \infty \tag{26}$$

Sketch of proof:
The proof boils down to proving the following relation

$$\sqrt{n}(K(\hat{F}_n) - K(F_0)) = \sqrt{n} \int \tilde{\theta}_{F_0}\, d(Q_n - Q_{F_0}) + o_p(1). \tag{27}$$

Then an application of the central limit theorem yields that the NPMLE of $K(F_0)$ has the desired asymptotically optimal behavior. The proof consists of the following steps.

I. By conditions (S1) and (D1), and lemma 4.1 we have

$$\sqrt{n}(K(\hat{F}_n) - K(F_0)) = \sqrt{n} \int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0) + o_p(1)$$

II. For $F \in \mathcal{F}$, one can define a function ϕ_F as a solution to the integral equation (19). This solution can be used to extend definition (20) to $\tilde{\theta}_F$ for $F \in \mathcal{F}$, where $\phi_F(u)/F(u)$ and $\phi_F(v)/(1 - F(v))$ are defined to be zero if $F(u) = 0$ or if $F(v) = 1$, respectively. Note that $\tilde{\theta}_F$ no longer has an interpretation as canonical gradient. In lemma 2.2 in GESKUS AND GROENEBOOM (1995B) the following is shown for $\tilde{\theta}_{\hat{F}_n}$:

$$\int \tilde{\kappa}_{F_0}\, d(\hat{F}_n - F_0) = -\int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0}.$$

III. Corollary 4.1 implies

$$\int \overline{\theta}_{\hat{F}_n}\, dQ_n = 0,$$

where $\overline{\theta}_{\hat{F}_n}$ denotes the function defined in (20), but with the function $\phi_{\hat{F}_n}$ replaced by $\overline{\phi}_{\hat{F}_n}$, which is constant on the intervals of constancy of the NPMLE (and equals $\phi_{\hat{F}_n}$ at one point of the interval). We then get

$$-\sqrt{n} \int \tilde{\theta}_{\hat{F}_n}\, dQ_{F_0} = \sqrt{n} \int \overline{\theta}_{\hat{F}_n}\, d(Q_n - Q_{F_0}) + \sqrt{n} \int (\overline{\theta}_{\hat{F}_n} - \tilde{\theta}_{\hat{F}_n})\, dQ_{F_0}$$

The second term can be shown to be $o_p(1)$.

IV. The first term is further split into

$$\begin{aligned}\sqrt{n}\int \bar{\theta}_{\hat{F}_n} d(Q_n - Q_{F_0}) &= \sqrt{n}\int \tilde{\theta}_{F_0} d(Q_n - Q_{F_0}) \\ &\quad + \sqrt{n}\int (\bar{\theta}_{\hat{F}_n} - \tilde{\theta}_{F_0}) d(Q_n - Q_{F_0})\end{aligned}$$

The last term can be shown to be $o_p(1)$, using a Donsker property of the class of functions under consideration.

References

AYER, M., BRUNK, H.D., EWING, G.M., REID, W.T., SILVERMAN, E. (1955). *An empirical distribution function for sampling with incomplete information*, Ann. Math. Statist., vol. 26, 641–647.

BAKKER, D. (1988). *Nonparametric maximum likelihood estimation of the distribution function of interval censored observations*, Master's thesis, University of Amsterdam.

BARLOW, R.E., BARTHOLOMEW, D.J., BREMNER, J.M., BRUNK, H.D. (1972). *Statistical Inference under Order Restrictions*, Wiley, New York.

BEGUN, J. M., HALL, W. J., HUANG, W. M., AND WELLNER, J. A. (1983). *Information and asymptotic efficiency in parametric - nonparametric models* Ann. Statist., vol. 11, 432–452.

BICKEL P.J., KLAASSEN C.A.J., RITOV Y. AND WELLNER J.A. (1993). *Efficient and adaptive estimation in semiparametric models*, John Hopkins University Press, Baltimore.

BIRMAN, M.Š., SOLOMJAK, M.Z. (1967). *Piecewise-polynomial approximations of functions in the classes W_p^α*. Math. Sbornik. vol. 73, 295–317.

DINSE, G. E. AND LAGAKOS, S. W. (1982). *Nonparametric estimation of lifetime and disease onset distributions from incomplete observations.* Biometrics, vol. 38, 921–932.

GEER S. VAN DE (1993). *Rates of convergence for the maximum likelihood estimator in mixture models*, Technical Report TW 93-09, University of Leiden.

GESKUS R.B. (1992). *Efficient estimation of the mean for interval censoring case II*, Technical Report 92-83, Delft University of Technology.

GESKUS R.B. AND GROENEBOOM P. (1995a). *Asymptotically optimal estimation of smooth functionals for interval censoring, part 1.* To appear in Statistica Neerlandica (jubilee issue).

GESKUS R.B. AND GROENEBOOM P. (1995b). *Asymptotically optimal estimation of smooth functionals for interval censoring, part 2.* Submitted to Statistica Neerlandica.

GESKUS R.B. AND GROENEBOOM P. (1995c). *Asymptotically optimal estimation of smooth functionals for interval censoring, case 2; observation times arbitrarily close*, Technical Report, Delft University of Technology, to appear.

GILL, R.D. AND LEVIT, B.Y. (1992) *Applications of the van Trees inequality: a Baysian Cramér-Rao bound.* Preprint Nr. 773, Department of Mathematics, University Utrecht.

GROENEBOOM, P. (1987). *Asymptotics for interval censored observations.* Technical Report 87-18, Department of Mathematics, University of Amsterdam.

GROENEBOOM, P. (1989). *Brownian motion with a parabolic drift and Airy functions.* Probability theory and related fields, vol. 81, 79–109.

GROENEBOOM, P. (1991). *Discussion on: Age-specific incidence and prevalence: a statistical perspective, by Niels Keiding.* J. R. Statist. Soc. A,vol. 154, 400–401.

GROENEBOOM P. AND WELLNER J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser Verlag.

HANSEN, B.E. (1991). *Nonparametric estimation of functionals for interval censored observations.* Master's thesis, Delft University of Technology and Copenhagen University.

HUANG J. AND WELLNER J.A. (1995a). *Asymptotic normality of the NPMLE of linear functionals for interval censored data, case 1*, to appear in Statistica Neerlandica.

HUANG J. AND WELLNER J.A. (1995b). *Efficient estimation for the proportional hazards model with "Case 2" interval censoring*, submitted.

JONGBLOED, G. (1995). *Three statistical inverse problems.* Ph.D. thesis, Delft University of Technology.

JONGBLOED G. (1995). *The iterative convex minorant algorithm for nonparametric estimation*, Technical Report, Delft University of Technology, to appear.

KEIDING, N. (1991) *Age-specific incidence and prevalence: a statistical perspective (with discussion).* J. R. Statist. Soc. A, vol. 154, 371–412.

KIM, J., POLLARD, D. (1990). *Cube root asymptotics.* Ann. Statist., vol. 18, 191–219.

KRESS R. (1989). *Linear integral equations*, Applied Mathematical Sciences vol. 82, Springer Verlag, New York.

PETO (1973). *Experimental survival curves for interval-censored data*, Appl. Statist. vol. 22, p. 86-91.

ROBERTSON, T., WRIGHT, F.T., DYKSTRA, R.L. (1988). *Order Restricted Statistical Inference.* Wiley, New York.

ROCKAFELLAR, R.T. (1970). *Convex analysis*, Princeton University Press.

SHEEHY, A. AND WELLNER, J. (1992). *Uniform Donsker Classes of Functions*, Ann. Prob.,vol 20, p. 1983-2030.

TURNBULL, B.W. (1974). *Nonparametric estimation of a survivorship function with doubly censored data.* J. Amer. Statist. Assoc., vol. 69, 169–173.

TURNBULL, B.W. (1976). *The empirical distribution function with arbitrarily grouped censored and truncated data.* J.R. Statist. Soc. B, vol. 38, 290–295.

TURNBULL, B. W. AND MITCHELL, T. J. (1984) *Nonparametric estimation of the distribution of time to onset for specific diseases in survival/sacrifice experiments.* Biometrics, vol. 40, 41–50.

VAN TREES H. L. (1968) *Detection, Estimation and Modulation Theory, Part 1.* Wiley, New York.

VAART A.W. VAN DER (1988). *Statistical estimation in large parameter spaces*, CWI Tract, vol. 44, Centrum voor Wiskunde en Informatica, Amsterdam.

VAART A.W. VAN DER (1991). *On differentiable functionals*, Ann. Statist., vol. 19, p. 178-204.