Nonparametric Estimation under Censoring and Passive Registration

Richard D. Gill

Mathematical Institute, University Utrecht,
Budapestlaan 6, 3584 CD Utrecht
Netherlands

gill@math.ruu.nl

Abstract. The classical random censorship model assumes that we follow an individual continuously up to the time of failure or censoring, so observing this time as well as the indicator of its type. Under passive registration we only get information on the state of the individual at random observation or registration times. In this paper we assume that these registration times are the times of events in an independent Poisson process, stopped at failure or censoring; the time of failure is also observed if not censored. This problem turns up in historical demography, where the survival time of interest is the life-length, censoring is by emigration, and the observation times are times of births of children, and other life-events. (Church registers contain dates of births, marriages, deaths, but not emigrations.) The model is shown to be related to the problem of estimating a density known to be monotone. This leads to an explicit description of the nonparametric maximum likelihood estimator of the survival function (based on i.i.d. observations from this model) and to an analysis of its large sample properties.

Key words and phrases. Nonparametric maximum likelihood estimator, exponential deconvolution, Grenander estimator.

Introduction.

Church registers contain dates of baptisms (roughly speaking: births), marriages, and burials (roughly: deaths); but not on immigrations and emigrations. In historical demography one uses church registers to estimate life-length and mobility in former centuries; see Blum (1989), Ruggles (1992). For individuals born in one village we observe a date of birth, followed by a sequence of dates of 'life-events' (marriage, births and sometimes also deaths of children, death of spouse, remarriage). For some individuals this sequence is terminated by the person's own death, the date of which is then also observed. However many emigrate away from the village during the course of their life. In that case the time of emigration is never observed, and nothing is known of what happened to the individual after that time. All we see are the life-events preceding emigration. By the absence of an observed death we infer that emigration took place some at completely unknown time after the last recorded life-event, and we know nothing else at all. We call this problem 'the passive registration problem' (J. Oeppen, Cambridge Group for the History of Population and Social Structure; personal communication).

Superficially this looks like a censoring problem. Considering the age at death as the survival time of interest, one might use the age at the last recorded life-event as a censoring time. However under any reasonable modelling, this would be incorrect. Individuals are at risk to die (their death being potentially observable) from the time of their last observed life-event right up to emigration. So if we disregard this fact we underestimate the number at risk at any time point, and overestimate the risk of death. The more emigration occurs, the stronger is this bias.

Before describing a formal statistical model for this problem, we specify the classical random censoring model for later comparison. According to the random censorship model, one observes n i.i.d. copies of the minimum $\widetilde{T}$ of a failure time T and an independent censoring time C together with an indicator $\Delta = 1\{T \leq C\}$ of the type of each observation. Thinking of the times T and C as being the times of two events (death or failure, and censoring respectively) in the life-time of an individual, this corresponds to continuous observation of the individual up to the time of the first occurring event.

In medical and biological applications with continuous monitoring of an individual, this may be a realistic model. In other fields it is however often unrealistic to assume continuous observation. Rather, the current status of an individual (failure already occurred/not yet occurred) is only observed or registered intermittently at some discrete time points, perhaps random and out of the control of the experimenter. For instance, in the interval censoring models studied in depth in Groeneboom and Wellner (1992), there are registration times R_i, independent of the survival time T, such that one only observes R_i and whether or not T is greater than R_i ($i = 1$ in 'case 1 interval censoring', $i = 1, 2$ in 'case 2').

Our passive registration model has features both from the random censoring model and the interval censoring model. We assume the three components survival (total life length), censoring (emigration), and intermittent registration (times of life-events) are independent. To be specific, T and C are independent times of death and censoring respectively, with unknown distributions. Independently of these, $R_1 < R_2 < \ldots$, are the times of events of some point process, taken throughout this paper to be a (not necessarily homogenous)

Poisson process. The data (from one individual) consists of the times R_i which precede $\widetilde{T} = \min(T, C)$ together with the time T if (and only if) $T < C$. The problem is to nonparametrically estimate the distribution F of T. We shall assume that the distribution G of C is also completely unknown but, in most of the paper, that the rate or the intensity function of the Poisson process is known. Assuming the rate is known, we may make a known time transformation to a unit rate Poisson process. The distributions of the transformed T and C remain completely unknown. After estimation on the basis of the transformed data we can transform back to the original time scale.

The data from one individual can be described equivalently as follows: we certainly observe $\Delta = 1\{T \leq C\}$. If $T \leq C$ we also observe the times R_i such that $R_i < T$ and T itself. If however $T > C$ we observe the times R_i such that $R_i < C$ but we observe neither C nor T.

Independence of the three processes (death, censoring, registration times) may seem far-fetched but it can be argued to be a pretty good first approximation. Definitely far-fetched is to assume that the registration times follow a Poisson process, and to assume that the intensity function of the process is known. However our aim is to analyse a tractable version of the problem in order to gain insight into the kind of phenomena which will be met with in non-parametric estimation of the survival function under passive registration, in more realistic models. Based on the succesful complete analysis of this special model, we can with confidence predict that the technique of non-parametric maximum likelihood estimation will also be succesful (though more complicated to implement) when applied to more realistic registration processes. Also we can predict important properties of the resulting estimators.

To discuss this we must first look at the more simple classical models. In both the random censorship model and the interval censoring models (case 1 or 2) nonparametric maximum likelihood estimators (NPMLEs) are appropriate and known to have various large sample optimality properties. However their asymptotic behaviour is quite different. Under random censorship, the NPMLE is the famous product-limit estimator of Kaplan and Meier (1958). It is consistent and converges at rate $n^{1/2}$ to a limiting Gaussian distribution about the true value. Under case 1 interval censoring however, the NPMLE converges at rate $n^{1/3}$ and the limiting distribution is non-Gaussian. The EM algorithm could in principle be used to compute the estimator, but in practice its convergence is too slow. Rather an algorithm related to isotonic regression (the derivative of the least concave majorant of a certain cumulative sum diagram) should be used. Under case 2 interval censoring the rate becomes $(n \log n)^{1/3}$ and the computation of the NPMLE more complicated (an iteratively reweighted version of the algorithm for case 1). Various functionals of the NPMLE in these delicate problems however have $n^{1/2}$, limiting Gaussian behaviour. Apart from the practical importance, these interesting mathematical phenomena, currently subject of much research, are a main motivation to study the passive registration problem.

In our problem the censoring time C is never observed and one might expect similar statistical properties (in particular, cube root of n asymptotics) of the NPMLE as in the interval censoring models just described. However we do have exact observations of the uncensored times, and this suggests root n asymptotics. We will see that under mild smoothness conditions

— the distribution function of T is estimated by the NPMLE at root n rate
— that of C at cube root of n rate;
these rates are optimal. (Smoothness is important: if the distribution of T is discrete while that of C is continuous, then the distribution of the former can only be estimated at cube root of n rate even though this seems to be a more 'parametric' model than with continuous, unknown distribution.) We expect these phenomena to be retained when the intensity of the Poisson process of registration times is not known, but is modelled parametrically; and even when we discard the Poisson assumption in favour of a more realistic renewal process model. We give some preliminary results in this direction.

We show that the NPMLE of the distribution of T and C can be computed directly through a variant of the isotonic regression method of the interval censoring models. The reason for this is that the Poisson process structure leads to an exponential deconvolution problem, which is known (Vardi, 1989; Groeneboom and Wellner, 1992) to be closely related to the Grenander problem of estimating a monotone density (Grenander, 1956); itself very close to the case 1 interval censoring model just mentioned. The EM algorithm can be expected to be very slow to converge; in fact, the more data is available, the slower, as is typically the case in non root n problems.

A direct approach to nonparametric estimation is to write down the likelihood and maximize it numerically. Alternatively, by considering the model as a missing data model (the censoring times of censored individuals are not observed), one can maximize the likelihood by means of the EM algorithm without ever actually looking at the likelihood. This comes down to using the Nelson-Aalen and Kaplan-Meier estimators (see Andersen, Borgan, Gill and Keiding, 1993) with the 'number at risk at time t' predicted by adding to the number definitely known to be at risk, the estimated probability that each censored individual has not yet emigrated. Either of these approaches might be necessary under more realistic modelling of the registration process, though one should always look for more effective algorithms. The careful analysis of our initial model reveals special structure which can be used to calculate the NPMLE more or less explicitly. This leads to a complete analysis of its statistical properties, which can be used to guess the properties of the NPMLE in more realistic models.

The model.

We consider the model underlying the observed data from one individual. Recall that $T \sim F$, $C \sim G$ and $R_1, R_2, \ldots$ are independent times of failure, censoring, and times of a unit rate Poisson process respectively. We observe the R_i with $R_i < \widetilde{T} = \min(T, C)$ and also T if $T \leq C$. Implicitly we also observe $\Delta = 1\{T \leq C\}$. However C itself is never observed, nor T if $T > C$.

Define T^* as the last time the individual is observed: thus $T^* = T$ if $\Delta = 1$, $T^* = \max\{R_i : R_i \leq C\}$ if $\Delta = 0$ where the maximum of the empty set is taken to be equal to 0.

Condition on T and C. One can imagine the times R_i with $R_i \leq \widetilde{T}$ as times of a unit rate Poisson process in reverse time, stopped at time 0. When $\Delta = 1$ this means that the conditional distribution of the observed R_i given the observed data $\widetilde{T}, \Delta$ is known (and

does not depend on F and G). So by sufficiency we may discard the registration times in that case.

Still conditioning, when $\Delta = 0$ the last registration time before $\widetilde{T}$ is distributed as the time of the first event in a unit rate Poisson process (in reverse time, starting at $\widetilde{T}$), except that, since the process is stopped at time 0, there may be no registration time at all. Recall that we defined T^* to be the last time the individual was observed. So in this case T^* has the same distribution as $\max(0, \widetilde{T} - E)$ where E denotes a unit exponential random variable. If $T^* > 0$ then the possible other registration times, taken in reverse order, are distributed as the times of a unit rate Poisson process starting at T^* and stopped at 0. Since this distribution also is fixed given the data T^*, Δ, by sufficiency they also may be discarded.

The conclusion of the above is that we may restrict attention, for each individual, just to the data T^*, Δ where T^* is the last time of observation of the individual and Δ indicates whether this was the time of failure T or a time of passive registration R_i (possibly equal to 0). Furthermore, $\Delta = 1\{T \leq C\}$ and if $\Delta = 1$ then $T^* = \widetilde{T}$, otherwise $T^* = \max(0, \widetilde{T} - E)$ where E is an independent unit exponential.

Write

$$F_i(t) = \Pr\{\widetilde{T} \leq t, \Delta = i\}, \quad i = 1, 0; \tag{1}$$

let $\widetilde{F} = F_1 + F_0$. We have

$$\begin{aligned} F_1(\mathrm{d}t) &= (1 - G(t-))F(\mathrm{d}t), \\ F_0(\mathrm{d}t) &= (1 - F(t))G(\mathrm{d}t). \end{aligned} \tag{2}$$

Moreover for any joint distribution (F_1, F_0) of a pair $(\widetilde{T}, \Delta)$ where $\widetilde{T}$ takes values in $[0, \infty)$ and Δ in $\{0, 1\}$, there exist (possibly defective) distributions F and G such that (1) holds; i.e., one can represent the distribution of *any* $(\widetilde{T}, \Delta)$ as the distribution of $\min(T, C), 1\{T \leq C\}$ for certain *independent* nonnegative T and C; one of them possibly taking the value $+\infty$ with positive probability; though since the distribution of $\widetilde{T}$ was supposed to be on $[0, \infty)$ not both of T and C can have positive probability to be infinite. Moreover the distributions of T and C are uniquely determined (at least) on $[0, \tau)$ where τ is the upper endpoint of the support of $\widetilde{T}$.

Briefly, letting Λ_F and Λ_G be the cumulative hazard functions of F and G, one reconstructs F and G from F_1 and F_0 through the relations $\mathrm{d}\Lambda_F = \mathrm{d}F_1/(1 - \widetilde{F}_-)$, $\mathrm{d}\Lambda_G = \mathrm{d}F_0/(1 - F_1 - F_{0-})$ and $1 - F = \prod(1 - \mathrm{d}\Lambda_F)$, $1 - G = \prod(1 - \mathrm{d}\Lambda_G)$; see Gill (1994).

In conclusion, as the distributions of T and C vary arbitrarily on $[0, \infty]$ (but not both with an atom at infinity), so does that of $(\widetilde{T}, \Delta)$ on $[0, \infty) \times \{0, 1\}$.

Our data is however (T^*, Δ). Writing

$$F_i^*(t) = \Pr\{T^* \leq t, \Delta = i\}, \quad i = 1, 0; \tag{3}$$

we see that as the distributions of T and C vary through all possible distributions on $[0, \infty]$ (not both with an atom at infinity), F_1^* and F_0^* vary through all pairs of subdistribution functions on $[0, \infty)$ whose sum is a distribution function on $[0, \infty)$ and such that F_0^* is

the convolution of an arbitrary subdistribution function with the distribution of minus a standard exponential, truncated at zero.

We therefore investigate the class of possible F_0^*. Since this characterization is of independent interest we first consider the case $T = \infty$ with probability one, so that $T^* = C$ and $\Delta = 0$ always; now $F_0 = G$.

Lemma. *Let $C \sim G$ on $[0,\infty)$ and let $C^* = \max(0, C - E)$ where E is standard exponential, independent of C. Then the distribution G^* of C^* has an atom g_0^* at zero but is absolutely continuous on $(0,\infty)$ with density g^* such that $e^{-t}g^*(t)$ is nonincreasing and (without loss of generality) right-continuous with left-hand limits. Moreover, $g_0^* \geq g^*(0)$. Conversely, any distribution G^* on $[0,\infty)$ with these properties can be uniquely represented as the distribution of $C^* = \max(0, C - E)$ for independent C and (unit exponential) E. If G lives on $(0,\infty)$ then $g_0^* = g^*(0)$.*

Proof. Starting with C, E we calculate, for $t > 0$, $G^*(\mathrm{d}t) = \int_{(t,\infty)} e^{-(s-t)}\mathrm{d}t G(\mathrm{d}s) = (e^t \int_{(t,\infty)} e^{-s} G(\mathrm{d}s))\mathrm{d}t = g^*(t)\mathrm{d}t$. So G^* has a density g^* on $(0,\infty)$ and $e^{-t}g^*(t)$ is decreasing, from its value $\int_{(0,\infty)} e^{-s}G(\mathrm{d}s)$ at $t = 0$ to zero at infinity. Moreover, $g_0^* = \int_{[0,\infty)} e^{-s}G(\mathrm{d}s) \geq g^*(0)$ with equality if and only if G has no atom at zero.

Conversely, let us suppose that G^* has all the given properties. Define, for $t > 0$, $G(\mathrm{d}t) = e^t(-\mathrm{d}(e^{-t}g^*(t)))$; let $G(\{0\}) = g_0^* - g^*(0)$. Because $e^{-t}g^*(t)$ is nonincreasing, G is a positive measure on $[0,\infty)$; with no mass at zero if $g_0^* = g^*(0)$. We have $G([0,t]) = g_0^* - g^*(0) + \int_{(0,t]} e^s(-\mathrm{d}(e^{-s}g^*(s))) = g_0^* - g^*(0) + \int_{(0,t]} e^s(e^{-s}g^*(s)\mathrm{d}s - e^{-s}\mathrm{d}g^*(s)) = g_0^* - g^*(0) + G^*(t) - G^*(0) - g^*(t) + g^*(0) = G^*(t) - g^*(t)$. Thus $G(t) = G([0,t])$ is nondecreasing, nonnegative and since $\liminf_{t\to\infty} g^*(t) = 0$ while $\limsup_{t\to\infty} G^*(t) = 1$ we must have $\lim_{t\to\infty} G(t) = \limsup_{t\to\infty}(G^*(t) - g^*(t)) = 1$; thus G is a distribution function. From the defining relation $G(\mathrm{d}t) = e^t(-\mathrm{d}(e^{-t}g^*(t)))$ one obtains $e^{-t}g^*(t) = \int_{(t,\infty)} e^{-s}G(\mathrm{d}s)$ or $g^*(t) = \int_{(t,\infty)} e^{-(s-t)}G(\mathrm{d}s)$; together with the other defining relation and $G(\{0\}) = g_0^* - g^*(0)$ this gives $g_0^* = G(0) + g^*(0) = \int_{[0,\infty)} e^{-s}G(\mathrm{d}s)$ which shows that G^* is the distribution of $\max(0, C - E)$ for certain C and E as described. □

Back to our model we now have: as F and G vary arbitrarily, the subdistributions F_i^*, $i = 1, 0$ also vary arbitrarily subject to their sum being a distribution function, and F_0^* having an atom p_0^* at zero and a density f_0^* on $(0,\infty)$ such that $e^{-t}f_0^*(t)$ is nondecreasing and (without loss of generality) right-continuous with left-hand limits; its limit for $t \to \infty$ is zero and for $t \to 0$ is less than or equal to p_0^*.

The proof of the lemma also shows how to reconstruct F_i, $i = 1, 0$ from F_i^*, $i = 1, 0$: of course $F_1 = F_1^*$, but F_0 is given by $F_0 = F_0^* - f_0^*$, or equivalently, $F_0(\mathrm{d}t) = e^t\mathrm{d}(-e^{-t}f_0^*(t))$, $t > 0$, $F_0(\{0\}) = p_0^* - f_0^*(0)$. This relation shows again that F_0 has an atom at zero if $p_0^* = F_0^*(0) > f_0^*(0)$. From the F_i we can reconstruct F and G (at least, on the support of $\widetilde{F} = F_0 + F_1$) via product integration of their hazard measures $\mathrm{d}\Lambda_F = \mathrm{d}F_1/(1 - \widetilde{F}_-)$, $\mathrm{d}\Lambda_G = \mathrm{d}F_1/(1 - F_1 - F_{0-})$.

One can further reparametrize the model through the probability $F_1^*(\infty) = 1 - F_0^*(\infty)$ to have an uncensored observation, and the conditional distributions $\mathrm{d}F_i^*/F_i^*(\infty)$, $i = 1, 0$, of T^* given $\Delta = i$. The probability and the first of the two probability distributions are

now completely arbitrary; the second is an arbitrary distribution of the type described by the lemma.

Derivation of the NPMLE.

By the characterization of the model for one observation, we see that the NPMLE of (F, G) based on n independent replications of (T^*, Δ) is obtained by estimating $F_1^*(\infty)$ by the fraction of uncensored observations; $F_1^*/F_1^*(\infty)$ is estimated by the empirical distribution function of the uncensored observations (with random sample size equal to their number); and $F_0^*/F_0^*(\infty)$ is estimated by the NPMLE of a distribution of the type described in the lemma, based on the censored observations (again, with random sample size equal to their number). We now describe the latter.

The lemma of the last section puts us into an exponential deconvolution model, with truncation. Now if E has the standard exponential distribution, then e^{-E} is uniformly distributed on $[0, 1]$. After exponential transformation, the operation of subtracting a standard exponential random variable becomes multiplication by a uniform $[0, 1]$. Vardi (1989) shows that the corresponding estimation problem (without truncation as in our case) is essentially the same as Grenander's (1956) problem of nonparametric estimation of a decreasing density; see also Groeneboom and Wellner (1992, Part II, Chapter 2, Exercise 3).

Suppose C^* has a distribution G^* of the type described in the lemma. Let $Y = e^{C^*}$, then the distribution H of Y has an atom of size g_0^* at $y = 1$ and a density h on $(1, \infty)$ equal to $(1/y)g_0^*(\log(y)) = (e^{-t}g^*(t))|_{t=\log y}$; i.e., $h(e^t) = e^{-t}g^*(t)$. So the distribution of Y has an atom at 1 and a nonincreasing, right-continuous density on $(1, \infty)$, bounded above by the size of the atom. We may compute the NPMLE of the distribution of C^*, given n i.i.d. observations, via that of $Y = e^{C^*}$. Denote the density of Y by f and the size of the atom by p_1. Then the NPMLE of (p_1, f) is obtained by maximizing $p_1^{\#\{i:Y_i=1\}} \prod_{i:Y_i>1} f(Y_i-)$; it is necessary to work with the left-continuous version of the density in the likelihood, otherwise the NPMLE does not exist. But if we let $\widetilde{f} = p_1$ on $[0, 1]$ and $\widetilde{f} = f_-$ on $(1, \infty)$, then $\widetilde{f}$ is a probability density (integrates to 1), nonincreasing and left-continuous, and constant on $[0, 1]$. Its NPMLE is the maximizer over such $\widetilde{f}$ of $\prod_i \widetilde{f}(Y_i)$. Now suppose we drop the requirement that $\widetilde{f}$ is constant on $[0, 1)$. Then the solution of the maximization problem over the larger class of $\widetilde{f}$ is the well-known Grenander estimate (Grenander, 1956) of a nonincreasing density (of a nonegative random variable): the left-hand derivative of the least concave majorant on $[0, \infty)$ of the empirical distribution function of the data Y_i. (One may check that this solution does not require the Y_i to be all different). The solution can be obtained as follows by the 'pool adjacent violators' algorithm. Consider the piecewise linear curve connecting the points $(0, 0)$ and $(Y_{(i)}, i/n)$, $i = 1, \ldots, n$, where $Y_{(i)}$ denote the order statistics of our sample. This plot is called the cumulative sum diagram. If the slopes of two adjacent line-segments are in the wrong order (the second one steeper than the first) then delete their joint endpoint from the diagram and replace the two line-segments by a single one. After a finite number of steps no more deletions are possible and we are done.

Note that since no Y_i are smaller than 1, the empirical distribution function makes its first jump, of size $\#\{i : Y_i = 1\}/n$, at $y = 1$. Its least concave majorant therefore has

constant slope over the interval $[0,1]$ of at least this size. It is possible that the slope then changes to a lower value; this happens if for all $Y_{(j)} > 1$, $\gamma_j = j/(nY_{(j)}) < \#\{i : Y_i = 1\}/n$. If however the maximum of the γ_j is larger than $\#\{i : Y_i = 1\}/n$, then the least concave majorant starts with the straight line connecting the origin to the corresponding $(Y_{(j)}, j/n)$.

But in any case, the maximizer over the larger class of nonincreasing $\widetilde{f}$ is actually a member of the smaller class of nonincreasing $\widetilde{f}$ constant on $[0,1]$. Therefore it also maximizes the likelihood over the smaller class and is the NPMLE we are looking for. To summarize: we compute the least concave majorant on $[0,\infty)$ of the empirical distribution function of the Y_i, and put $\widehat{p}_1$ equal to its (constant) slope on $[0,1)$, and $\widehat{f}$ (which we choose to be right continuous) equal to its right hand derivative on $[1,\infty)$.

Description of the NPMLE.

Now we can put all the above ingredients together to describe the NPMLE $\widehat{F}_i$ of $F_i, i = 1, 0$. A note on notation: a hat indicates a maximum likelihood estimator, a superscript (n) indicates an empirical distribution function. We simply let $\widehat{F}_1$ be the empirical subdistribution function $F_1^{(n)}$ of the T_i^* with $\Delta_i = 1$. Compute the least concave majorant $\widehat{H}_0$ of the subdistribution function $H_0^{(n)}$ of the $e^{T_i^*}$ with $\Delta_i = 0$. Let $\widehat{h}_0$ be its *right-continuous* (sub)density, which is constant on $[0,1)$. Then we estimate F_0 by $\widehat{F}_0$ defined for $t \geq 0$ by $\widehat{F}_0(t) = \widehat{H}_0(e^t) - e^t\widehat{h}_0(e^t)$. Equivalently, $\widehat{F}_0$ has an atom at 0 of size $\widehat{H}_0(1) - \widehat{h}_0(1) = -(\widehat{h}_0(1) - \widehat{h}_0(1-))$ while for $t > 0$, $\widehat{F}_0(\mathrm{d}t) = e^t(-\mathrm{d}(\widehat{h}_0(e^t)))$; in fact this can be combined to give $\widehat{F}_0(\mathrm{d}t) = e^t(-\mathrm{d}(\widehat{h}_0(e^t)))$ on $t \geq 0$.

Finally we estimate F and G by the usual product-integration of estimated hazards. Let $\widehat{H}_1 = H_1^{(n)}$ denote the empirical subdistribution function of the $e^{T_i^*}$ with $\Delta_i = 1$. Then

$$\begin{aligned}
\widehat{\Lambda}_F(t) &= \int_{[0,t]} \frac{\widehat{F}_1(\mathrm{d}s)}{1 - \widehat{F}_1(s-) - \widehat{F}_0(s-)} \\
&= \int_{[0,t]} \frac{\widehat{F}_1(\mathrm{d}s)}{1 - \widehat{F}_1(s-) - \widehat{H}_0(e^s-) + e^s\widehat{h}_0(e^s-)} \\
&= \int_{[1,e^t]} \frac{\widehat{H}_1(\mathrm{d}y)}{1 - \widehat{H}_1(y-) - \widehat{H}_0(y-) + y\widehat{h}_0(y-)}
\end{aligned} \tag{4}$$

and

$$1 - \widehat{F}(t) = \prod_{[0,t]} (1 - \widehat{\Lambda}_F(\mathrm{d}s)).$$

Similarly,

$$\begin{aligned}\widehat{\Lambda}_G(t) &= \int_{[0,t]} \frac{\widehat{F}_0(\mathrm{d}s)}{1-\widehat{F}_1(s)-\widehat{F}_0(s-)} \\ &= \int_{[0,t]} \frac{\widehat{F}_0(\mathrm{d}s)}{1-\widehat{F}_1(s)-\widehat{H}_0(e^s-)+e^s\widehat{h}_0(e^s-)} \\ &= \int_{[1,e^t]} \frac{-y\widehat{h}_0(\mathrm{d}y)}{1-\widehat{H}_1(y)-\widehat{H}_0(y-)+y\widehat{h}_0(y-)} \qquad (5)\end{aligned}$$

and

$$1-\widehat{G}(t) = \prod_{[0,t]}(1-\widehat{\Lambda}_G(\mathrm{d}s)).$$

Our description of $\widehat{F}_0$ shows that it is quite possible for $\widehat{F}_0$ and hence $\widehat{G}$ to have an atom at $t=0$ even if the true G has support on $(0,\infty)$: this happens if $\widehat{H}_0$ has a kink (change of slope) at $y=1$. If we had looked for the NPMLE in the smaller class of F and G without atom at zero, it would have been possible for the NPMLE not to exist. This is the reason we were careful to develop the model and the estimators allowing for atoms at zero. The results of Woodroofe and Sun (1993) actually show that if G has no atom at zero, then the probability tends to 1 as $n\to\infty$ that $\widehat{H}_0$ has no kink at $y=1$. So the probability does tend to one that the NPMLE within the smaller class of distributions on $(0,\infty)$ exists, if the truth lies in this class too.

Large sample theory: overview.

We begin with some general comments on the estimation of F. Denote, as before, by $H_i^{(n)}$ the empirical sub-distribution functions of the $e^{T_j^*}$ with $\Delta_j = i$, $i=1,0$. So $\widehat{H}_1 = H_1^{(n)}$ but $\widehat{H}_0$ is the least concave majorant of $H_0^{(n)}$ on $[0,\infty)$. In expression (4) we see entering: the empirical distribution function $H_1^{(n)}$ and the least concave majorant $\widehat{H}_0$ together with its density $\widehat{h}_0$ on $(1,\infty)$. Now $n^{1/2}(\widehat{H}_1 - H_1)$ is an ordinary empirical process and converges in distribution to a Gaussian limit. Under the minimal assumption of strict concavity (of H_0 on $[1,\infty)$) it will be shown that $n^{1/2}(\widehat{H}_0 - H_0)$ and $n^{1/2}(H_0^{(n)} - H_0)$ are asymptotically equivalent on $[1,\infty)$ (their difference converges in distribution to the zero process). Under further smoothness and strictness conditions (continuous differentiability of h_0 with a strictly negative derivative) the finite dimensional distributions of $n^{1/3}(\widehat{h}_0 - h_0)$ converge to nondegenerate, non-Gaussian limits (with independent coordinates). This makes it not so easy to see what the asymptotic behaviour of $\widehat{\Lambda}_F$ is. In fact the result depends crucially on the smoothness of F_1, and thereby on the smoothness of F.

Suppose F is actually a discrete distribution. Then $\widehat{\Lambda}_F$ is a sum over the jump-times of F and its asymptotic behaviour will be dominated by the cube root of n behaviour of $\widehat{h}_0$ at these jump-times.

Suppose on the other hand that F has a density which is of bounded variation. We prove in the technical appendix that a linearization or first order Taylor expansion of

(4) about its limiting value, the same expression with the hats removed, is valid. We sketch how this linearization, together with asymptotic equivalence of $n^{1/2}(\widehat{H}_0 - H_0)$ and $n^{1/2}(H_0^{(n)} - H_0)$, leads to root n behaviour of $\widehat{\Lambda}_F$.

To cut down the notation, we drop the range and variable of integration in (4) and use a subscript $-$ to denote the left continuous versions of the functions in the denominator of (4). The identity function $y \mapsto y$ also appearing in the denominator of (4) is denoted by ι. An integral without limits denotes the function $x \mapsto \int_{[1,x]} \ldots$. If we assume F has a density, then H_1 and $\widehat{H}_1$ live on $(1,\infty)$ rather than on $[1,\infty)$, and we may take the integrals over $(1,x]$ instead of $[1,x]$. In either case we may rewrite (4) minus its supposed limit, evaluated at $t = \log x$, as

$$\begin{aligned}(\widehat{\Lambda}_F - \Lambda_F) \circ \log &= \int \frac{\mathrm{d}\widehat{H}_1}{1 - \widehat{H}_{1-} - \widehat{H}_{0-} + \iota\widehat{h}_{0-}} - \int \frac{\mathrm{d}H_1}{1 - H_{1-} - H_{0-} + \iota h_{0-}} \\ &\approx \int \frac{\mathrm{d}(\widehat{H}_1 - H_1)}{1 - H_{1-} - H_{0-} + \iota h_{0-}} \\ &\quad + \int \frac{(\widehat{H}_{1-} - H_{1-})\mathrm{d}H_1}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2} \\ &\quad + \int \frac{(\widehat{H}_{0-} - H_{0-})\mathrm{d}H_1}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2} \\ &\quad - \int \frac{\iota(\widehat{h}_{0-} - h_{0-})\mathrm{d}H_1}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2}.\end{aligned}$$

The first three terms converge at root n rate to asymptotically Gaussian limits, being integrals of, or with respect to, empirical processes (or at least, processes asymptotically equivalent to empirical processes). If H_1 has a density h_1, we may rewrite the fourth term as

$$-\int \frac{\iota(\widehat{h}_{0-} - h_{0-})\mathrm{d}H_1}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2} = -\int \frac{\iota h_1 \mathrm{d}(\widehat{H}_0 - H_0)}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2}.$$

Here we are integrating *with respect to* a process asymptotically equivalent to an empirical process. Provided the integrand, and in particular the density h_1, is of bounded variation, integration by parts expresses this as an integral *of* a process asymptotically equivalent to an empirical process. The final hoped for result is then:

$$\begin{aligned}(\widehat{\Lambda}_F - \Lambda_F) \circ \log \approx &\int \frac{\mathrm{d}(H_1^{(n)} - H_1)}{1 - H_{1-} - H_{0-} + \iota h_{0-}} \\ &+ \int \frac{(H_{1-}^{(n)} - H_{1-})\mathrm{d}H_1}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2} \\ &+ \int \frac{(H_{0-}^{(n)} - H_{0-})\mathrm{d}H_1}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2} \\ &- \int \frac{\iota h_1 \mathrm{d}(H_0^{(n)} - H_0)}{(1 - H_{1-} - H_{0-} + \iota h_{0-})^2}.\end{aligned} \tag{6}$$

We show in the appendix that (6) is true, with $\approx$ interpreted to mean that the difference between left and right hand sides is $o_P(n^{-1/2})$ in supremum norm on any closed interval $[1, \tau]$ on which the denominators in (6), $1 - H_{1-} - H_{0-} + \iota h_{0-}$ or its square, are bounded away from zero. Precise conditions for this to hold are, on each such interval: h_1 is of bounded variation, and h_0 is continuous and strictly decreasing. The use of the first condition has just been demonstrated, while the latter condition makes $\widehat{h}_0$ uniformly consistent, which is all we need to know about it when carrying out the linearization. It also makes H_0 strictly concave, which we need for the earlier mentioned asymptotic equivalence. In terms of the underlying F and G the conditions are, on each interval $[0, \tau]$ such that $F(\tau) < 1$, $G(\tau) < 1$, that F has a density f of bounded variation, and G is continuous and strictly increasing. Here we also make use of the fact that the function appearing in the denominators of (6) is nothing else than the left-continuous version of $(1-F)(1-G) \circ \log$.

The consequence of (6) with this interpretation of $\approx$ is that $n^{1/2}(\widehat{\Lambda}_F - \Lambda_F)$ converges in distribution to the same limit as that of root n times the right hand side of (6); this being a standard (function indexed) empirical process, converging in distribution to a Gaussian limit. Since the product-integral mapping taking hazard to distribution is sufficiently smooth, this carries over to convergence in distribution of $n^{1/2}(\widehat{F} - F)$ (by the functional delta method, see for instance Gill, 1994, Section 6).

For estimating G the situation is quite different. Linearization of (5) and then integration by parts leads to terms having square root of n behaviour together with a term equal to some function of e^t times $\widehat{h}_0(e^t) - h_0(e^t)$. If H_0 is two times continuously differentiable with second derivative bounded away from zero, this term has cube root of n behaviour, and will dominate the others. So $n^{1/3}(\widehat{\Lambda}_G - \Lambda_G)$ converges pointwise in distribution to a scaled version of the limiting distribution of $\widehat{h}_0$, and the same will hold for $n^{1/3}(\widehat{G} - G)$.

More general registration processes.

We now discuss what can be done when the registration process does not have a completely known distribution. As a first step we suppose it is a Poisson process, with unknown but constant rate α.

If the rate α was actually known, one could make the time transformation $s = \alpha t$ to a unit rate process, then estimate the corresponding transformed F and G by our previously described methods, then finally transform back. For given (known) α, the estimates $\widehat{F}$ and $\widehat{G}$ so obtained are the NPMLE's of F, G.

Since this transformation, but using an estimate of α instead of the true but unknown value, will play a role in the discussion below, we describe it in more detail. Let F_α and G_α be defined by $F_\alpha(s) = \mathrm{P}(\alpha T \leq s) = F(s/\alpha)$; $G_\alpha(s) = \mathrm{P}(\alpha C \leq s) = G(s/\alpha)$. Replace the observations T^*, Δ by $\alpha T^*, \Delta$. Estimate F_α and G_α by the procedure of the previous section applied to these transformed observations and then transform back using $F(t) = F_\alpha(\alpha t)$, $G(t) = G_\alpha(\alpha t)$.

When α is unknown, two procedures for estimating F, G, and α come quickly to mind. A fairly simple, ad hoc, procedure is based on the fact that conditional on T^*, Δ, the registration times $R_i < T^*$ are times in a Poisson process of rate α observed on the time interval $[0, T^*]$. The *conditional maximum likelihood estimator* $\overline{\alpha}$ of α based on this part of the data is simply the total, over the n observations, of the number of registration

times strictly before each observed T^* divided by the total of the observed values of T^*. Now carry out the transformation procedure to estimate F and G using just the observed values of T^*, Δ, pretending that $\alpha = \overline{\alpha}$.

A more sophisticated but more respectable procedure is to use joint non-parametric maximum likelihood. Since the NPMLE of F, G for given α is easy to find, it is natural to use *profile likelihood* to estimate α. Let $\overline{N}$ denote the mean observed number of registration times (including T^* itself if $\Delta = 0$), let $\overline{\Delta T^*}$ denote the mean of the n observed values of ΔT^*. With this notation the ad hoc estimator of α described above is $\overline{\alpha} = \overline{N - 1\{\Delta = 0, T^* > 0\}}/\overline{T^*}$. Let $H_{0,\alpha}$ denote the subdistribution function and $h_{0,\alpha}$ the density of $e^{\alpha T^*}$ with $\Delta = 0$. Let $H_{0,\alpha}^{(n)}$ denote the corresponding empirical subdistribution function, $\widehat{H}_{0,\alpha}$ denote its least concave majorant, and $\widehat{h}_{0,\alpha}$ the left-hand derivative thereof. The subdensity f_0^* of the T^* with $\Delta = 0$ is easily written down in terms of $h_{0,\alpha}$ and α for any given α. We may parametrize by α, $h_{0,\alpha}$, and $F_1^* = F_1$, the subdistribution of the T^* with $\Delta = 1$. The likelihood for these three parameters factors into a $(\alpha, h_{0,\alpha})$ part and an F_1^* part; the latter resulting in the corresponding empirical as NPMLE. So we look further only at the $(\alpha, h_{0,\alpha})$ part.

Some routine calculations show that $1/n$ times the log likelihood for $(\alpha, h_{0,\alpha})$ can be written as

$$\overline{N} \log \alpha - \overline{\Delta T^*}\alpha + \int_{[1,\infty)} \log h_{0,\alpha} \mathrm{d}H_{0,\alpha}^{(n)}.$$

Therefore $1/n$ times the profile likelihood for α is

$$\begin{aligned} \overline{N} \log \alpha - \overline{\Delta T^*}\alpha + \int_{[1,\infty)} \log \widehat{h}_{0,\alpha} \mathrm{d}H_{0,\alpha}^{(n)} \\ = \overline{N} \log \alpha - \overline{\Delta T^*}\alpha + \int_{[1,\infty)} \widehat{h}_{0,\alpha} \log \widehat{h}_{0,\alpha}. \end{aligned}$$

It seems very feasible to compute this profile likelihood for a grid of values of α in a suitable neighbourhood of $\overline{\alpha}$ and maximize it numerically.

Preliminary investigations show that the asymptotic properties of the ad hoc procedure can be derived on very similar lines to our analysis in the fixed α case. The main technical problem is the question of weak convergence of the empirical process based on the transformed observations $\overline{\alpha} T^*$, where $\overline{\alpha}$ is dependent of all the observations and itself asymptotically normally distributed. This can be done using the functional delta-method and the differentiability of the composition operator, see Andersen, Borgan, Gill and Keiding (1993, Section II.8 and especially Proposition II.8.8). The ad hoc estimators are well behaved and in particular $\overline{\alpha}$ and the estimator of F are asymptotically jointly normal at root n rate; it is not too difficult to write down the precise asymptotic distribution.

The NPMLE's are harder to study, and so far we did not obtain complete results. If the maximum likelihood estimator of α could be shown to be root n consistent, the further analysis would be straightforward and use just the tools which have been developed above. Aad van der Vaart (personal communication) has been able to establish this result.

It is plausible that the joint NPMLE of (α, F) is actually asymptotically normal and moreover asymptotically equivalent to the ad hoc estimators. If well-behaved at all, it can

be expected to be asymptotically efficient; see Gill and van der Vaart (1993). We have in fact been able to prove by analysis of the so-called tangent space for our model (Bickel, Klaassen, Ritov and Wellner, 1993), that the ad hoc estimator is actually asymptotically efficient in the semi-parametric sense.

For practical purposes, what is important is that the ad hoc estimator is actually a very sensible estimator to use. Apparently the extra information about α hidden in the observations of T^* with $\Delta = 0$ is not of importance compared to the information in the conditional distribution of the registration times given T^*. This means that the obvious extension of the ad hoc procedure to the case when the registration process is modelled by an inhomogenous Poisson process, depending perhaps on several unknown parameters, is not only easy to carry out but also asymptotically efficient.

The NPMLE needs further study. When the registration process is no longer modelled by a Poisson process but, for example, a renewal process, analogues of our ad hoc procedure are no longer available. However NPMLE seems at least computationally feasible and we may expect that its statistical properties are good too. More urgently needed than heavy theoretical results is practical experience with realistic modelling of passive registration data. The mathematical analysis we have carried out here suggests that non-parametric maximum likelihood estimation will be a useful and reliable tool.

Bibliography.

P.K. Andersen, Ø. Borgan, R.D. Gill, and N. Keiding (1993), *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore.

P. Billingsley (1968), *Convergence of Probability Measures*, Wiley, New York.

A. Blum (1989), An estimate of local adult mortality based on family cards, *Population* **44**, English Selection No. 1, 39–56.

R.D. Gill (1994), Lectures on survival analysis; in: D. Bakry, R.D. Gill and S.A. Molchanov, *Lectures on Probability Theory, (École d'Éte de Probabilités de Saint-Flour XXII–1992)*, ed. P. Bernard, Springer Lecture Notes in Mathematics **1581**.

R.D. Gill and A.W. van der Vaart (1993), Non- and semi-parametric maximum likelihood estimators and the von Mises method, Part II, *Scandinavian Journal of Statistics* **20**, 271–288.

R.D. Gill, M.J. van der Laan and J.A. Wellner (1993), Inefficient estimators of the bivariate survival function for three multivariate models, Preprint 767, Dept. Math., Univ. Utrecht; to appear in *Annales de l'Institute Henri Poincaré*.

U. Grenander (1956), On the theory of mortality measurement, Part II, *Skandinavisk Aktuarietidskrift* **39**, 125–153.

P. Groeneboom (1985), Estimating a monotone density, pp. 539–555, in: L. Le Cam and R.A. Olshen (eds), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II*, Wadsworth, Monterey.

P. Groeneboom (1989), Brownian motion with a parabolic drift and Airy functions, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **81**, 79–109.

P. Groenboom and H.P. Lopuhäa (1993), Isotonic estimators of monotone densities and distribution functions: basic facts, *Statistica Neerlandica* **47**, 175–184.

P. Groeneboom and J.A. Wellner (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, DMV Seminar vol. 19, Birkhäuser, Basel.

J. Huang and J.A. Wellner (1993), Estimation of a monotone density or monotone hazard under random censoring, Tech. Rep. 252, Dept. Statist., Univ. Washington; to appear in *Scandinavian Journal of Statistics.*

E.L. Kaplan and P. Meier (1958), Non-parametric estimation from incomplete observations, *Journal of the Amererican Statistical Association* **53**, 457–481, 562–563.

J. Kiefer and J. Wolfowitz (1976), Asymptotically minimax estimation of concave and convex distribution functions, *Zeitschrift für Wahrsheinlichkeitstheorie und verwandte Gebiete* **34**, 73–85.

J. Kim and D. Pollard (1990), Cube root asymptotics, *Annals of Statistics* **18**, 191–219.

D. Pollard (1984), *Convergence of Stochastic Processes*, Springer-Verlag, Berlin.

D. Pollard (1990), *Empirical Processes: Theory and Applications*, Regional conference series in probability and statistics **2**, Institute of Mathematical Statistics, Hayward, California.

B.L.S. Prakasa Rao, Estimation of a unimodal density, *Sankhya (Series A)* **31**, 23–36.

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

S. Ruggles (1992), Migration, marriage and mortality: correcting sources of bias in English family reconstitutions, *Population Studies* **46**, 507–522.

A.W. van der Vaart and J.A. Wellner (1995), *Weak Convergence and Empirical Processes*, to appear.

Y. Vardi (1989), Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation, *Biometrika* **76**, 751–761.

M. Woodroofe and J. Sun (1993), A penalised maximum likelihood estimate of $f(0+)$ when f is non-increasing, *Statistica Sinica* **3**, 501–515.

Appendix: technical lemmas for large sample theory.

To fill in the sketch of the asymptotic behaviour of $\widehat{F}$ we need two kinds of technical result. Firstly, we must justify the statement that $\sqrt{n}(\widehat{H}_0 - H_0^{(n)})$ converges in supremum norm, on appropriate intervals, in probability to zero; here, $H_0^{(n)}$ is an empirical sub-distribution function and $\widehat{H}_0$ its least concave majorant, the underlying H_0 is strictly concave on $[1, \infty)$, zero on $[0, 1)$, and $H_0(1) \geq h_0(1)$. We call this the asymptotic equivalence problem for the least concave majorant. Secondly we must take a careful look at the linearization leading from (4) to (6). We need solutions to both problems which work when the underlying Poisson process depends on unknown parameters which have to be estimated simultaneously with F and G. This results in replacement of the transformed observations $e^{T_i^*}$ by something more complicated in which the transformation also depends on an estimated parameter. So we will not be able to use results for i.i.d. observations. We will just make use of the fact that $\sqrt{n}(H_0^{(n)} - H_0)$ and $\sqrt{n}(H_1^{(n)} - H_1)$ converge jointly in distribution to some limiting process.

The second problem (linearization) is quite simple to solve, using standard techniques (telescoping, integration by parts, and the Helly-Bray technique; see e.g., Gill, van der Laan and Wellner, 1993). We return to this later. The first problem (asymptotic equivalence of least concave majorant) has been first studied by Kiefer and Wolfowitz (1976), using many specific properties of the empirical distribution function. By their deep analysis stronger results are obtained than we need, but not in a general enough context. Similar comments can be made on the related results given by Huang and Wellner (1993). Further key references on the Grenander estimator are Prakasa Rao (1969), Groeneboom (1985, 1989), Robertson, Wright and Dykstra (1988), Kim and Pollard (1990), Groeneboom and Lopuhäa (1993), Woodroofe and Sun (1993).

First we look at the problem shorn of the special features concerning the interval $[0, 1]$.

Theorem 1. *Let F be a bounded, nondecreasing, strictly concave function on $[0, \infty)$ and let F_n be an estimator of F, which is a right-continuous, nondecreasing step function. Suppose F is continuous at zero and its right-hand derivative there is finite. Let $\widehat{F}$ denote the least concave majorant of F_n on $[0, \infty)$. Suppose that F_n converges in supremum norm on $[0, \infty)$ to F, in probability, and that $Z_n = \sqrt{n}(F_n - F)$ converges in distribution to some limiting process Z in $D[0, \infty)$, under the supremum norm on compact intervals, in the sense of Pollard (1984) (with respect to the open-ball sigma algebra). Suppose Z has continuous sample paths almost surely. Then $\sqrt{n}(\widehat{F} - F_n)$ converges in probability in the supremum norm on each compact interval to zero.*

Note 1. If the reader prefers, weak convergence may be understood in the more modern sense (with respect to the Borel sigma-algebra, but using outer expectation), see Pollard (1990) or van der Vaart and Wellner (1995), or in the classical (Billingsley, 1968) sense.

Note 2. The interval $[0, \infty)$ does not play any special role in the proof and could be replaced by any other interval throughout. What is relevant is that its right-hand end-point is not included. Typically in applications some of our conditions break down at the endpoint; moreover the result of the theorem typically cannot be extended to the closed interval $[0, \infty]$.

Note 3. Since F is concave, it is continuous and has finite right-hand and left-hand derivatives everywhere, except possibly at the end-point $t = 0$. Our assumption extends these properties also to $t = 0$.

Proof. By a Skorohod-Dudley almost sure construction the whole sequence F_n and also the limiting process Z can be considered as defined on a single probability space, having the original marginal distributions, but satisfying now $Z_n = \sqrt{n}(F_n - F)$ converges almost surely in supremum norm on compact intervals to Z, and F_n converges almost surely in supremum norm on the whole interval to F. It suffices to show that for this representation, $\sqrt{n}(\widehat{F} - F_n)$ converges almost surely in supremum norm on compact intervals to 0. Then we have the corresponding convergence in probability for the original objects. Fix $\sigma < \tau < \infty$ and write $\widehat{F}^\tau$ for the least concave majorant of F_n on the interval $[0, \tau]$. We show that

(i) $\sqrt{n}(\widehat{F}^\tau - F_n) = o(1)$ almost surely with respect to the supremum norm on $[0, \tau]$,

(ii) $\widehat{F}$ and $\widehat{F}^\tau$ coincide on $[0, \sigma]$ almost surely for all large enough n.

These facts give the required result for the interval $[0, \sigma]$.

To start with fact (ii), almost surely, for large enough n and given δ, F_n lies between $F \pm \delta$. We assumed that F is strictly concave: this implies that for given $\lambda \in (0,1)$, $F(\lambda\sigma + (1-\lambda)\tau) > \lambda F(\sigma) + (1-\lambda)F(\tau)$. For small enough δ therefore

$$F(\lambda\sigma + (1-\lambda)\tau) - \delta > \lambda(F(\sigma)+\delta) + (1-\lambda)(F(\tau)+\delta). \tag{7}$$

Now if F_n lies within δ of F, $F+\delta$ is *a* concave majorant of F_n while $\widehat{F}$ is *the least* concave majorant so $F+\delta \geq \widehat{F} \geq F_n \geq F-\delta$. Therefore $\widehat{F}$ lies above $F-\delta$ at $\lambda\sigma+(1-\lambda)\tau$ but below $F+\delta$ at σ and at τ. Because F_n is a right-continuous, non-decreasing step function, $\widehat{F}$ is piecewise linear with kinks (changes of derivative) at certain jump points of F_n. By (7) $\widehat{F}$ must have a kink in $[\sigma,\tau]$, so there exists t in this interval where $\widehat{F}(t) = \widehat{F}_n(t)$. The least concave majorant on $[0,\infty)$, restricted to $[0,\tau]$, is a concave majorant on $[0,\tau]$ so this shows that also $\widehat{F}^\tau(t) = F_n(t)$. It can be shown from this that $\widehat{F}$ and $\widehat{F}^\tau$ coincide to the left of t and hence in particular on the interval $[0,\sigma]$. (Draw a picture, or consider the concave majorant of F_n on $[0,\infty)$ equal to $\widehat{F}^\tau$ on $[0,t]$ extended linearly with slope equal to the left-hand derivative of $\widehat{F}$ at t on (t,∞).)

Now for fact (i). For the rest of the proof we work only on the interval $[0,\tau]$. We write $\widehat{F}$, also l.c.m.(F_n), for the least concave majorant of F_n *on this interval* and $\|\cdot\|$ for the supremum norm on the same interval. The following argument applies to almost all realisations and we omit the otherwise many times repeated statement 'almost surely'. We let f denote the right-hand derivative of F; it is nonincreasing, right-continuous with left-hand limits, and finite, and for all $t \neq t_0$ we have by strict concavity of F that $F(t) < F(t_0) + (t-t_0)f(t_0)$. In other words, the straight line $t \mapsto F(t_0) + (t-t_0)f(t_0)$ lies above the graph of F touching it at $t = t_0$ only. The same applies to the straight line obtained by replacing $f(t_0)$ by $f(t_0-)$.

If $\|Z - Z_n\| \leq \varepsilon$ then $\|(F + n^{-1/2}Z_n) - (F + n^{-1/2}Z)\| \leq n^{-1/2}\varepsilon$. So the least concave majorant of $F + n^{-1/2}Z_n$, plus $n^{-1/2}\varepsilon$, is a concave majorant of $F + n^{-1/2}Z$, and vice-versa. Thus $n^{1/2}\|\text{l.c.m.}(F + n^{-1/2}Z_n) - \text{l.c.m.}(F + n^{-1/2}Z)\| \leq \varepsilon$; so we have $n^{1/2}\|\text{l.c.m.}(F + n^{-1/2}Z_n) - \text{l.c.m.}(F + n^{-1/2}Z)\| \to 0$ as $n \to \infty$. It suffices therefore to show $\|n^{1/2}(\text{l.c.m.}(F + n^{-1/2}Z) - F) - Z\| \to 0$ as $n \to \infty$.

Now $n^{1/2}(\text{l.c.m.}(F+n^{-1/2}Z)-F)-Z \geq n^{1/2}((F+n^{-1/2}Z)-F)-Z = 0$ so it suffices to show that

$$\limsup_{n\to\infty} \sup_{[0,\tau]} (n^{\frac{1}{2}}(\text{l.c.m.}(F + n^{-\frac{1}{2}}Z) - F) - Z) = 0.$$

Moreover, the operation 'least concave majorant' commutes with addition of a linear function, and is dominated by 'supremum', so

$$\begin{aligned} & n^{\frac{1}{2}}(\text{l.c.m.}(F + n^{-1/2}Z)(t_0) - F(t_0)) - Z(t_0) \\ & \qquad = n^{\frac{1}{2}}\text{l.c.m.}\Big(F + n^{-1/2}Z - F(t_0) - n^{-\frac{1}{2}}Z(t_0) - (\cdot - t_0)(f(t_0) + n^{-\frac{1}{2}}c)\Big)(t_0) \\ & \qquad \leq n^{\frac{1}{2}} \sup\Big(F + n^{-\frac{1}{2}}Z - F(t_0) - n^{-\frac{1}{2}}Z(t_0) - (\cdot - t_0)(f(t_0) + n^{-\frac{1}{2}}c)\Big). \end{aligned}$$

Here we subtracted the linear function, zero at $t = t_0$, equal to $t \mapsto (t-t_0)(f(t_0)+n^{-1/2}c)$, where c is arbitrary. In the present proof we can take $c = 0$, but when we later modify the argument to take account of different assumptions concerning the interval $[0,1]$, it will be necessary to take another choice of c. But now we have

$$\sup_{[0,\tau]}(n^{\frac{1}{2}}(\text{l.c.m.}(F + n^{-\frac{1}{2}}Z) - F) - Z)$$
$$\leq \sup_{t_0,t\in[0,\tau]} \Big(Z(t) - Z(t_0) + (t-t_0)c + n^{\frac{1}{2}}\big(F(t) - F(t_0) - (t-t_0)f(t_0)\big)\Big)$$

Suppose the lim sup as $n \to \infty$ of the last displayed quantity is positive. Then we can find t_{0n} and t_n such that the limit of the following quantity, along some subsequence n_k, exists and is positive:

$$\Big(Z(t_n) - Z(t_{0n}) + (t_n - t_{0n})c\Big) + n^{\frac{1}{2}}\Big(F(t_n) - F(t_{0n}) - (t_n - t_{0n})f(t_{0n})\Big). \qquad (8)$$

Pick a further subsequence along which t_n and t_{0n} both converge to say t and t_0; and if necessary by picking a further subsequence arrange that t_{0n} approaches t_0 from one side.

On the finally chosen subsequence $F(t_n) - F(t_{0n}) - (t_n - t_{0n})f(t_{0n})$ converges to $F(t) - F(t_0) - (t-t_0)f(t_0\pm)$ which is strictly negative unless $t = t_0$. We also have by continuity of Z that $Z(t_n) - Z(t_{0n}) + (t_n - t_{0n})c$ converges to the finite quantity $Z(t) - Z(t_0) + (t-t_0)c$. If $t \neq t_0$ then the limit of (8) is $-\infty$; if $t = t_0$ then it is nonpositive, so in either case we have a contradiction. □

Note 4. The Kiefer and Wolfowitz (1976) version of this result assumes twice continuous differentiability of F and uses delicate empirical process results, on the other hand a rate of convergence is also obtained. Groeneboom (1994, personal comunication) has a short proof of the pointwise result but again using twice continuous differentiability.

Note 5. If F is not strictly concave while Z is not constant on an interval where F is linear, the theorem fails.

We now give a version of the theorem for the case of interest to us. The previous proof still works, except that we have to be careful in choosing c in the final part of the argument.

Theorem 2. *Let F and F_n be defined on $[0,\infty)$ as in theorem 1, but both are zero on $[0,1)$ and make a jump upwards at $t = 1$. F is bounded, nondecreasing; strictly concave on $[1,\infty)$; and F_n is a right-continuous, nondecreasing step function. Suppose the right-hand derivative f of F is finite at $t = 1$ and satisfies $F(1) \geq f(1)$. Let $\widehat{F}$ denote the least concave majorant of F_n on $[0,\infty)$. Suppose that F_n converges in supremum norm on $[0,\infty)$ to F, in probability, and that $Z_n = \sqrt{n}(F_n - F)$ converges in distribution to some limiting process Z in $D[0,\infty)$, under the supremum norm on compact intervals. Note that Z_n and Z are identically zero on $[0,1)$. Suppose Z has continuous sample paths on $[1,\infty)$ almost surely;*

however we do not assume $Z(1) = 0$. *Then* $\sqrt{n}(\widehat{F} - F_n)$ *converges in probability in the supremum norm on each compact subinterval of* $[1, \infty)$ *to zero.*

Proof. The first part of the previous proof (establishing fact (ii)) goes through without change for $\tau, \sigma > 1$, which is sufficient for our purposes. We next, in analogy to fact (i), want to show asymptotic equivalence on intervals of the form $[1, \tau]$, but with the least concave majorants computed relative to $[0, \tau]$. It makes no difference then to redefine F, F_n, Z and Z_n on $(0, 1)$ by their linear interpolants between the points 0 and 1. This makes F and Z continuous on $[0, \infty)$. Following the previous line of proof leads us to consider

$$\limsup_{n \to \infty} \sup_{t_0 \geq 1, t \geq 0} \Big(Z(t) - Z(t_0) + (t - t_0)c + n^{\frac{1}{2}}(F(t) - F(t_0) - (t - t_0)f(t_0)) \Big).$$

It is now important to choose the value of c carefully: we take $c = -Z(1)$. Let us look separately at the cases $t \geq 1$ and $t < 1$. The case $t \geq 1$ needs no alteration. For the case $t < 1$, with $c = -Z(1)$, we have

$$\begin{aligned} Z(t) - Z(t_0) + (t - t_0)c + n^{\frac{1}{2}}(F(t) - F(t_0) - (t - t_0)f(t_0)) \\ = tZ(1) - Z(t_0) - (t - t_0)Z(1) + n^{1/2}(tF(1) - F(t_0) - (t - t_0)f(t_0)). \end{aligned}$$

Now by strict concavity on $[1, \tau]$, and the fact that $F(1) \geq f(1)$, we have for $t \leq 1, t_0 \geq 1$ that $tF(1) \leq F(t_0) + (t - t_0)f(t_0)$ with equality only if $t_0 = 1$ and $F(1) = f(1)$. With $t_0 = 1$ the term $tZ(1) - Z(t_0) - (t - t_0)Z(1)$ equals zero, whatever the value of $t \leq 1$. The previous argument by convergent subsequences therefore works again. For suppose that, along a subsequence, t_n and t_{0n} asymptotically achieve the $\limsup_n \sup_{t_0, t}$ and converge (in the case of t_{0n} from one side) to certain t and t_0 respectively. If $t_0 > 1$, the term $n^{1/2}(F(t_n) - F(t_{0n}) - (t_n - t_{0n})f(t_{0n}))$ converges to $-\infty$ while the term $t_n Z(1) - Z(t_{0n}) - (t_n - t_{0n})Z(1)$ converges to something finite. If however $t_0 = 1$, then the first term, while not necessarily diverging, remains nonpositive; the second term converges to zero. In both cases the limit cannot be positive. □

Note 6. Continuing in the case when everything is zero on $[0, 1)$ but jumps at $t = 1$, suppose that actually f is continuous and $f(1) = F(1)$. Since F_n converges uniformly on $[0, \infty)$ in probability to F, it is easy to check that $\widehat{F}$ converges uniformly in probability to F modified by interpolating linearly between 0 and 1. Note that

$$(\widehat{F}(t) - \widehat{F}(t - \varepsilon))/\varepsilon \geq \widehat{f}(t) \geq (\widehat{F}(t + \varepsilon) - \widehat{F}(t))/\varepsilon$$

for all t and all $\varepsilon > 0$. Since the outer sides of these inequalities converge in probability to quantities arbitrarily close to $f(t)$ if ε is sufficiently small (here we use continuity of f) we see that $\widehat{f}$ is pointwise consistent; since it is monotone and the limit is continuous, it is uniformly consistent. Another useful fact is that we have shown that $n^{1/2}(\widehat{F}(1) - F_n(1))$ converges in probability to zero, though in general as we remarked earlier we will have $\widehat{F}(1) \neq F_n(1)$. In the classical Grenander problem (estimating a density f monotone on $[0, \infty)$) it has been shown by Woodroofe and Sun (1993) that $\widehat{f}(0)$ is inconsistent;

they propose a penalized maximum likelihood approach to solve this. It seems promising alternatively to adapt our solution, and to estimate f by a sieved maximum likelihood estimator: estimate f subject to f is constant on $[0, \varepsilon_n]$ where $\varepsilon_n \to 0$ as $n \to \infty$ is chosen suitably (a topic for future research).

Now we turn to the linearization problem. Recall that the aim is to show weak convergence of $n^{1/2}$ times the process (defined by taking the integral sign as short-hand for the mapping $x \mapsto \int_{(1,x]}$)

$$\int \frac{\mathrm{d}\widehat{H}_1}{1 - \widehat{H}_{1-} - \widehat{H}_{0-} + \iota\widehat{h}_{0-}} - \int \frac{\mathrm{d}H_1}{1 - H_{1-} - H_{0-} + \iota h_{0-}}. \tag{9}$$

We work on a closed interval $[1, \tau]$ on which $1 - H_{1-} - H_{0-} - \iota h_{0-}$ is bounded away from zero. We know that H_1 and $H_0 - \iota h_0$ are sub-distribution functions, adding to a distribution function; H_0 has density h_0 which is nonincreasing, right-continuous with left-hand limits. The same statements can be made about the estimated quantities.

We assume h_0 is actually continuous and strictly decreasing so that (among other things) H_0 is strictly concave. We assume that H_1 has a density h_1 which is of bounded variation.

We have that the empirical processes $n^{1/2}(H_1^{(n)} - H_1)$ and $n^{1/2}(H_0^{(n)} - H_0)$ converge jointly in distribution to a pair of Brownian bridge type processes on $[1, \tau]$, with respect to the supremum norm. By continuity of H_0 and H_1 the limiting process has continuous sample paths. By Theorem 2 (for the second component) the processes $n^{1/2}(\widehat{H}_1 - H_1)$ and $n^{1/2}(\widehat{H}_0 - H_0)$ have the same limit. By Theorem 2 and Note 6, the estimators $\widehat{H}_1$, $\widehat{H}_0$ and $\widehat{h}_0$ are monotone and converge uniformly on $[1, \tau]$ in probability to the quantities they are estimating.

We now apply telescoping which means, in (9), subtracting and adding intermediate terms in which the 'hats' are removed first from the numerator and then from the denominator. After that we rewrite expressions like $h_0 \mathrm{d}H_1$ as $h_1 \mathrm{d}H_0$, just as we did in obtaining (6). The difference (9) can then be written as a sum of four integrals:

$$\int \frac{\mathrm{d}\widehat{H}_1}{1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-}} - \int \frac{\mathrm{d}H_1}{1-H_{1-}-H_{0-}+\iota h_{0-}}$$

$$= \int \frac{\mathrm{d}(\widehat{H}_1 - H_1)}{1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-}}$$

$$+ \int \frac{(\widehat{H}_{1-} - H_{1-})\mathrm{d}H_1}{(1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-})(1-H_{1-}-H_{0-}+\iota h_{0-})}$$

$$+ \int \frac{(\widehat{H}_{0-} - H_{0-})\mathrm{d}H_1}{(1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-})(1-H_{1-}-H_{0-}+\iota h_{0-})}$$

$$- \int \frac{\iota(\widehat{h}_{0-} - h_{0-})\mathrm{d}H_1}{(1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-})(1-H_{1-}-H_{0-}+\iota h_{0-})}$$

$$= \int \frac{\mathrm{d}(\widehat{H}_1 - H_1)}{1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-}}$$

$$+ \int \frac{(\widehat{H}_{1-} - H_{1-})\mathrm{d}H_1}{(1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-})(1-H_{1-}-H_{0-}+\iota h_{0-})}$$

$$+ \int \frac{(\widehat{H}_{0-} - H_{0-})\mathrm{d}H_1}{(1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-})(1-H_{1-}-H_{0-}+\iota h_{0-})}$$

$$- \int \frac{\iota h_1 \mathrm{d}(\widehat{H}_{0-} - H_{0-})}{(1-\widehat{H}_{1-}-\widehat{H}_{0-}+\iota\widehat{h}_{0-})(1-H_{1-}-H_{0-}+\iota h_{0-})}.$$

Multiply throughout by $n^{1/2}$. Each of the four integrals now is an integral of, or with respect to, the weakly converging processes $n^{1/2}(\widehat{H}_1 - H_1)$ and $n^{1/2}(\widehat{H}_0 - H_0)$. The remaining parts of the integrands converge uniformly, in probability, to deterministic functions, and moreover their variation is bounded in probability. Now the Helly-Bray technique used in Gill (1989, Lemma 3), or see Gill, van der Laan and Wellner (1993, Lemma 2.5), gives us weak convergence of the integrals to their natural limits. Moreover, since the processes $n^{1/2}(\widehat{H}_1 - H_1)$ and $n^{1/2}(\widehat{H}_0 - H_0)$ are asymptotically equivalent to the empirical processes $n^{1/2}(H_1^{(n)} - H_1)$ and $n^{1/2}(H_0^{(n)} - H_0)$, the limiting distribution of the process $n^{1/2}(\widehat{\Lambda}_F - \Lambda_F)$ is indeed the one described by (6).