Bernoulli **7**(1), 2001, 1–31

A semi-parametric model for censored and passively registered data

MARIANNE A. JONKER* and AAD W. VAN DER VAART**

*Free University Amsterdam, Division of Mathematics and Computer Science, Free University, De Boelelaan 1081a, 1081 HV Amsterdam, Netherlands. E-mail: *majonker@cs.vu.nl; **aad@cs.vu.nl*

We consider the estimation of the parameters in a semi-parametric model for life-history data from historical demography. The data consist of a sequence of times of life events that is either ended by a time of death or right-censored by an unobserved time of migration. We derive the properties of the maximum likelihood estimators of the parameters and prove their asymptotic efficiency. Estimating the migration distribution turns out to be an inverse problem, whereas the other parameters are regular. The proof is based on a uniform rate of convergence of the Grenander estimator of a monotone density and bounds on the number and spacings of its support points.

Keywords: Grenander estimator; maximum likelihood; nuisance parameter; survival analysis

Introduction

To estimate life length and mobility in England from the sixteenth to the eighteenth century historical demographers use data from parish registers. Dates of baptisms (births), marriages and burials (deaths) were routinely recorded in parish registers, but in many parishes the registers are incomplete or individuals who are listed cannot be identified. Due to mobility, data concerning one person may be scattered over several registers and, despite much effort, it is not always possible to link the records. As a consequence, the times of death of approximately 40% of all people are missing. In these cases we observe the time of birth and possibly the times of a sequence of life events: marriage, births and deaths of children, death of husband and remarriage. The main reason why a time of death is missing is thought to be emigration to another parish. Unfortunately, the time of emigration is not observed, so that the sequence of life events is right-censored at an unobserved censoring time. See Wrigley and Schofield (1983) and Wrigley *et al*. (1997) for extensive discussions of data of this type.

Since the censoring time is unobserved, the data cannot be handled by standard techniques for censored survival data. For this reason Gill (1997) introduces a 'passive registration' type of censoring. In this model the observations are derived from three independent processes. The first is the time of death T, the second the time of censoring C and the third is the process $R_0 \equiv 0, R_1, R_2, \ldots$ of 'registration events'. The observations concerning one person consist of

$$\Delta = 1\{T \leq C\}, \qquad \Delta T, \qquad R_1, R_2, \ldots, R_N,$$

where

1350–7265

$$N = \max\{n \geqslant 0 : R_n \leqslant T \wedge C\}.$$

Note that we never observe the 'time of migration' C and observe the 'time of death' T only if death takes place before migration. Gill (1997) assumes that the registration events $R_1, R_2, \ldots$ are the events of a Poisson process with known intensity and studies the asymptotic properties of the maximum likelihood estimators of the distributions F and G of T and C, respectively, under the assumption that these distributions are completely unknown. In this paper we drop the assumption of a known rate and study the maximum likelihood estimator of the rate θ, F and G jointly. In the course of our proofs we establish a number of results on the Grenander estimator of a monotone density that are of independent interest.

For ease of notation, define T^* to be the last moment at which a person is seen to be alive, i.e.

$$T^* = \begin{cases} T & \text{if } \Delta = 1, \\ R_N, & \text{if } \Delta = 0. \end{cases}$$

It is convenient to reparametrize the model in terms of θ and the subdistribution functions

$$F^*_{0,\theta}(t) = \mathrm{P}_{\theta,F,G}(T^* \leqslant t, \Delta = 0),$$

$$F^*_1(t) = \mathrm{P}_{F,G}(T^* \leqslant t, \Delta = 1).$$

(Both $F^*_{0,\theta}$ and F^*_1 depend on F and G, but we do not let this show up in the notation.) Straightforward calculations, for instance using properties of the Poisson process, show that the density for a single observation $X = (\Delta, \Delta T, R_1, \ldots, R_N)$ can be written in the form

$$\Delta\theta^N \mathrm{e}^{-\theta T^*} f^*_1(T^*) + (1-\Delta)1\{N > 0\}\theta^{N-1}\mathrm{e}^{-\theta T^*} f^*_{0,\theta}(T^*) + (1-\Delta)1\{N = 0\}F^*_{0,\theta}\{0\}.$$

Here $f^*_{0,\theta}$ and f^*_1 are densities of the respective subdistribution functions and $F^*_{0,\theta}\{0\}$ denotes a point mass at zero. It is shown below that $F^*_{0,\theta}$ is absolutely continuous on $(0, \infty)$ and its (Lebesgue) density $f^*_{0,\theta}$ can be taken to be left-continuous; we shall use this version to define a likelihood. If we write $F^*_{0,\theta}\{0\}$ as $f^*_{0,\theta}(0)$, then $f^*_{0,\theta}$ is a density relative to the sum of the Dirac measure at 0 and Lebesgue measure on $(0, \infty)$ and the preceding display can be abbreviated to

$$\theta^{N-1\{\Delta=0,T^*>0\}}\mathrm{e}^{-\theta T^*} f^*_1(T^*)^{\Delta} f^*_{0,\theta}(T^*)^{1-\Delta}. \tag{1.1}$$

The density f^*_1 will be seen to be arbitrary; to define a likelihood we shall replace the term $f_1(T^*)$ by the point mass $F^*_1\{T\}$ and thus create a mixed empirical and ordinary likelihood. Throughout we assume that the total set of observations is a random sample $X_1, \ldots, X_n$ from the distribution of X and define the total likelihood as the product over the observations of the likelihoods for the n individuals.

In the expression (1.1) for the likelihood the original parameters are hidden in the distributions $F^*_{0,\theta}$ and F^*_1, but they can be recovered by explicit formulae. First, we define the subdistribution functions

$$F_0(t) = \mathrm{P}_{F,G}(T \wedge C \leqslant t, \Delta = 0) = \int_{[0,t]} (1 - F)(s)\, \mathrm{d}G(s),$$

$$F_1(t) = \mathrm{P}_{F,G}(T \wedge C \leqslant t, \Delta = 1) = \int_{[0,t]} (1 - G)(s-)\, \mathrm{d}F(s).$$

It is known (see Gill 1994) that the pair (F_0, F_1) ranges over all pairs of defective distribution functions on $[0, \infty)$ that add up to a distribution function as (F, G) ranges over all pairs of distribution functions on $[0, \infty]$ such that at least one of F and G concentrate on $(0, \infty)$. Furthermore, there exists a one-to-one relationship between the restrictions of the pair (F, G) and the pair (F_0, F_1) to the interval where $1 - F$ and $1 - G$ are positive. In fact, the preceding display can be explicitly inverted through the product integrals (see Gill 1994)

$$\begin{aligned} 1 - F(t) &= \prod_{0 \leqslant s \leqslant t} (1 - \Lambda_F\{s\}) \mathrm{e}^{-\Lambda_F^c(t)}, \qquad & \mathrm{d}\Lambda_F &= \frac{\mathrm{d}F_1}{1 - F_{0-} - F_{1-}}, \\ 1 - G(t) &= \prod_{0 \leqslant s \leqslant t} (1 - \Lambda_G\{s\}) \mathrm{e}^{-\Lambda_G^c(t)}, \qquad & \mathrm{d}\Lambda_G &= \frac{\mathrm{d}\,F_0}{1 - F_1 - F_{0-}}, \end{aligned} \tag{1.2}$$

where the superscript c denotes the continuous part and a minus sign denotes a left-continuous version. Thus (θ, F, G) can be recovered from (θ, F_0, F_1). Second, the triples $(\theta, F_{0,\theta}^*, F_1^*)$ and (θ, F_0, F_1) possess the relationships (for $t > 0$)

$$F_0(t) = F_{0,\theta}^*(t) - \frac{1}{\theta} f_{0,\theta}^*(t+), \qquad F_1 = F_1^*. \tag{1.3}$$

Here the second relation is obvious and the first relation follows from the following lemma, adapted from Gill (1997), which also characterizes the possible values of the new parameter $F_{0,\theta}^*$.

The lemma concerns the distribution of T^* given that $\Delta = 0$. If $\Delta = 0$, then $T^* = R_N$ is the last event before $T \wedge C = C$ of a Poisson process that is started independently of (T, C) at zero. Thus T^* is conditionally distributed as $\max(0, C - E)$ for a variable E that is independent of C and has an exponential distribution with rate θ. Given $\Delta = 0$, the variable $C = T \wedge C$ has distribution function $\tilde{F}_0 = F_0 / F_0(\infty)$.

Lemma 1.1. *Let $\tilde{F}_{0,\theta}^*$ be the distribution of* $\max(0, C - E)$*, where C is a random variable with a distribution function $\tilde{F}_0$ on* $[0, \infty]$ *and E is independent of C and has an exponential distribution with rate θ. Then $\tilde{F}_{0,\theta}^*$ has an atom $\tilde{F}_{0,\theta}^*\{0\}$ at zero, and is absolutely continuous on* $(0, \infty)$ *with a density $\tilde{f}_{0,\theta}^*$ that can be chosen such that the function $t \mapsto \mathrm{e}^{-\theta t} f_{0,\theta}^*(t)$ is non-increasing, left-continuous and such that $\tilde{f}_{0,\theta}^*(0+) \leqslant \theta \tilde{F}_{0,\theta}^*\{0\}$. Conversely, any distribution $\tilde{F}_{0,\theta}^*$ on* $[0, \infty]$ *with these properties can be uniquely represented as the distribution of* $\max(0, C - E)$ *for C and E as given, where $\tilde{F}_0$ can be recovered as $\tilde{F}_0(t) = \tilde{F}_{0,\theta}^*(t) - \theta^{-1} \tilde{f}_{0,\theta}^*(t+)$. Finally, we have $\tilde{f}_{0,\theta}^*(0+) = \theta \tilde{F}_{0,\theta}^*\{0\}$ if and only if $\tilde{F}_0\{0\} = 0$.*

Up to the (unknown) constant $F_0(\infty) = \mathrm{P}(\Delta = 0)$, the preceding lemma gives the set of possible distributions $F_{0,\theta}^*$ over which we must maximize the likelihood. By the preceding

discussion, the parameters $F_1^* = F_1$ and $F_{0,\theta}^*$ are connected only through the requirement that the total masses $F_1(\infty)$ and $F_{0,\theta}^*(\infty)$ add up to 1, but otherwise vary independently, where $F_1^* = F_1$ can be any subdistribution function on $[0, \infty]$.

To carry out the maximization it is convenient to rewrite the likelihood a second time. Define the subdistribution function

$$H_{0,\theta}(t) = \mathrm{P}_{\theta,F,G}(\mathrm{e}^{\theta T^*} \leqslant t, \Delta = 0).$$

The corresponding subdistribution has an atom $H_{0,\theta}\{1\} = F_{0,\theta}^*\{0\}$ at 1 and a density $h_{0,\theta}$ on $(1, \infty)$ such that

$$t \mapsto h_{0,\theta}(\mathrm{e}^{\theta t}) = \frac{1}{\theta}\mathrm{e}^{-\theta t} f_{0,\theta}^*(t) = \int_{[t,\infty)} \mathrm{e}^{-s\theta}\,\mathrm{d}F_0(s) \tag{1.4}$$

is non-increasing and such that $H_{0,\theta}\{1\} \geqslant h_{0,\theta}(1+)$, in view of the preceding lemma (with equality if and only if $F_0\{0\} = 0$). The first is equivalent to $h_{0,\theta} : (1, \infty) \mapsto \mathbb{R}$ being non-increasing. If we write the point mass at 1 as $h_{0,\theta}(1)$, then the two requirements can be described together by saying that $h_{0,\theta} : [1, \infty) \mapsto \mathbb{R}$ is non-increasing, and the likelihood can be rewritten in the form

$$\mathrm{l}(\theta, F, G)(X) = \theta^N \mathrm{e}^{-\theta T^* \Delta} F_1^*\{T^*\}^\Delta h_{0,\theta}(\mathrm{e}^{\theta T^*})^{1-\Delta}. \tag{1.5}$$

The following lemma, slightly adapted from Gill (1997), shows how to compute the maximum likelihood estimators of F_1^* and $h_{0,\theta}$, if θ is known.

Lemma 1.2. *If $\theta > 0$ is known, then the maximum likelihood estimator for F_1 is given by*

$$\hat{F}_1(t) = \frac{1}{n}\sum_{i=1}^n 1\{T_i \leqslant t, \Delta_i = 1\}.$$

Furthermore, the maximum likelihood estimator for $h_{0,\theta}$ is the left derivative of the least concave majorant of the subdistribution function

$$\mathbb{H}_{0,\theta}(t) := \frac{1}{n}\sum_{i=1}^n 1\{\mathrm{e}^{\theta T_i^*} \leqslant t, \Delta_i = 0\}.$$

Proof. The parameters F_1 and $h_{0,\theta}$ are only connected through the requirement that $F_1(\infty) + H_{0,\theta}(\infty) = 1$. We can release this connection by introducing an additional parameter $p = F_1(\infty)$ and replacing F_1 and $H_{0,\theta}$ in the likelihood by $p\tilde{F}_1$ and $(1-p)\tilde{H}_{0,\theta}$, where $\tilde{F}_1$ and $\tilde{H}_{0,\theta}$ range independently over parameter sets as before but are also restricted to be distribution functions. This yields the likelihood

$$\prod_{i=1}^n \theta^{N_i}\mathrm{e}^{-\theta T_i \Delta_i}\tilde{F}_1\{T_i\}^{\Delta_i} p^{\Delta_i}(1-p)^{1-\Delta_i}\tilde{h}_{0,\theta}(\mathrm{e}^{\theta T_i^*})^{1-\Delta_i}.$$

Maximizing this over p readily yields $\hat{p} = n^{-1}\sum_{i=1}^n \Delta_i$ as the maximum likelihood estimator. Furthermore, maximizing over $\tilde{F}_1$ yields $\sum_{i=1}^n 1\{T_i \leqslant t, \Delta_i = 1\}/\sum_{i=1}^n 1\{\Delta_i = 1\}$. The first assertion of the lemma is now immediate.

To prove the second assertion, we can identify $\tilde{H}_{0,\theta}$ with the absolutely continuous distribution on $(0, \infty)$ that is identical to $\tilde{H}_{0,\theta}$ on $(1, \infty)$ and has a density that is identically equal to $\tilde{H}_{0,\theta}\{1\}$ on $(0, 1]$. (Thus we spread the point mass at 1 uniformly over the interval $(0, 1)$.) Then the functions $t \to \tilde{h}_{0,\theta}(t)$ range exactly over all non-increasing probability densities on $(0, \infty)$ that are constant on $(0, 1]$. With $Y_1, \ldots, Y_k$ denoting the values $e^{\theta T_i^*}$ for which $\Delta_i = 0$, the maximization of the likelihood over $\tilde{h}_{0,\theta}$ is precisely the maximization of $\prod_i h(Y_i)$ over the set of all monotone densities h on $(0, \infty)$ that are constant on $(0, 1]$. Since $Y_i \geqslant 1$ for every i, the requirement that h be constant on $(0, 1]$ is not operational and hence the maximization yields the Grenander estimator. This is the left derivative of the least concave majorant of the empirical distribution function $\mathbb{H}_{0,\theta}/(1 - \hat{p})$ of $Y_1, \ldots, Y_k$. (See Robertson *et al.* 1988; Groeneboom and Lopuhaa 1993; or van der Vaart 1998, Section 24.4.) □

The maximum likelihood estimators for the pair $(H_{0,\theta}, F_1)$, can be transformed into maximum likelihood estimators for (F, G), as explained previously, for every fixed θ. To find the true maximum likelihood estimators we form the profile likelihood for θ by reinserting the maximum likelihood estimators for $(H_{0,\theta}, F_1)$ for known θ, maximize this over θ by a grid search or a Newton algorithm, and finally reinsert the maximum likelihood estimator for θ. We have implemented this procedure on computer and show a picture of the profile log-likelihood function in Figure 1.

Instead of the maximum likelihood estimator, there is a different, simpler estimator of θ, namely

$$\hat{\theta}_{n,c} := \frac{\sum_{i=1}^{n} (N_i - 1\{T_i^* > 0, \Delta_i = 0\})}{\sum_{i=1}^{n} T_i^*}. \tag{1.6}$$

This is the conditional maximum likelihood estimator of θ based on the total number of registration events in the interval $[0, T^*)$ given T^* (and $T^* > 0$). We show below that this is also the solution of the 'efficient score equation', from which it readily follows that this estimator is asymptotically efficient. An alternative to using the maximum likelihood estimator for the full parameter (θ, F, G) would be to use the *ad hoc* estimator for θ and then the estimators for F and G resulting from the preceding lemma.

The main results of this paper are the asymptotic distributions of the maximum likelihood estimators for θ, F and G. The maximum likelihood estimators for θ and F converge at rate $\sqrt{n}$ to Gaussian distributions and are efficient in the semi-parametric sense. A slight modification of the maximum likelihood estimator for G converges at rate $n^{-1/3}$ to a non-Gaussian limit.

In our proofs of these results we use properties of the Grenander estimator, or rather of the modification of this estimator described in Lemma 1.2. In Section 8 we show, *inter alia*, that the uniform rate of the Grenander estimator is of order $(n/\log n)^{-1/3}$ and that the spacings between these support points are of order $(n/\log n)^{-1/3}$.

The model as discussed in this paper can be viewed as a first attempt to tackle the

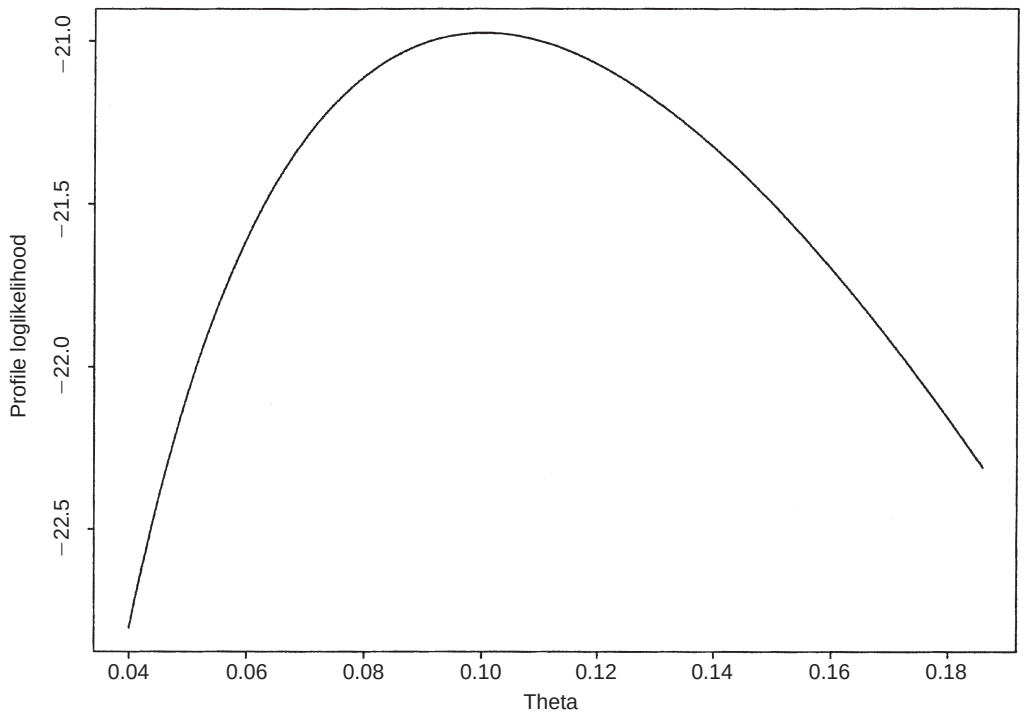

Figure 1. Profile log-likelihood for the estimation of θ based on a sample of size 5000. The true value of θ is 0.1.

historical life-length data mentioned above. The model has the benefit of being mathematically tractable, albeit that the analysis is already involved. A number of changes may help to make the model more realistic for this particular data set. For instance, it appears realistic to make the censoring time (interpreted as a time of moving) dependent on certain events in the registration process (such as marriage). Furthermore, the Poisson character of the registration process may not be fully realistic. Alternative models, still with a similar structure to the model in this paper, are studied by Jonker (2000), who also discusses the implementation of estimating procedures for these models and proves their consistency.

The remainder of the paper is organized as follows. In Section 2 we establish the consistency of the maximum likelihood estimator by Wald's method. This is standard, but a necessary step in the proof of the main results. In Section 3 we derive the tangent space of the model and compute the efficient score function for θ. The form of the latter score function is an important motivation for the proof of the asymptotic normality of the maximum likelihood estimator for θ, given in Section 4. Next, Sections 5 and 6 contain the results concerning the estimation of F and G, respectively. The details of the proofs of these results are given in Sections 9 and 10. Finally, Sections 7 and 8 are appendices which contain further details, and the results on the Grenander estimator mentioned previously.

Throughout the paper we assume that the true distributions F and G are continuous and are supported inside a bounded subinterval of $[0, \infty)$.

2. Consistency

In this section we show that the maximum likelihood estimators for θ, F and G are consistent on the intervals where they are identifiable. This is important as a first step in the derivation of the asymptotic distributions. The proof is an application of the method of Wald (1949), after first eliminating F_1 from the likelihood.

Theorem 2.1. *For every $\tau > 0$ such that $(1 - F)(\tau)(1 - G)(\tau) > 0$, the maximum likelihood estimators satisfy $(\hat{\theta}, \hat{F}, \hat{G}) \xrightarrow{\mathrm{p}} (\theta, F, G)$ for the product of the Euclidean distance and twice the uniform distance on $[0, \tau]$.*

Proof. As in the proof of Lemma 1.2, the likelihood for one observation can be written in the form

$$\mathrm{l}(\theta, p, \tilde{F}_0, \tilde{F}_1)(X) = \theta^N \mathrm{e}^{-\theta T^* \Delta} \tilde{F}_1\{T\}^\Delta p^\Delta (1 - p)^{1-\Delta} \left(\int_{[R_N, \infty)} \mathrm{e}^{-\theta s} \, \mathrm{d}\tilde{F}_0(s) \right)^{1-\Delta}.$$

The maximum likelihood estimator for $\tilde{F}_1$ was explicitly found in Lemma 1.2 and is seen to be uniformly consistent by the Glivenko–Cantelli theorem. Thus we can drop the corresponding term and study the likelihood as a function of $(\theta, p, \tilde{F}_0)$ only. To prove consistency we apply Wald's general consistency theorem (see Wald 1949; or van der Vaart 1998, Section 5.2.1). This has three main conditions: continuity of the likelihood in the parameters; integrability of the (local) suprema of the log-likelihood ratios; and identifiability.

Wald's method works best if the parameter set is compact. We choose the parameter sets for θ, p, and $\tilde{F}_0$ respectively equal to $[0, \infty]$, $[0, 1]$, and the set of distribution functions on $[0, \tau_0]$ equipped with the topology of weak convergence. Here τ_0 is an upper bound on the support of R_N. Then it is necessary to extend the definition of the likelihood to the case where $\theta \in \{0, \infty\}$, which we perform by continuity (ignoring the case where $(T^*, \Delta) = (0, 1)$, which has probability zero):

$$\mathrm{l}(0, p, \tilde{F}_0)(X) = \begin{cases} 0 & \text{if } N > 0, \\ p^\Delta (1 - p)^{1-\Delta} \displaystyle\int_{[0,\infty)} \mathrm{d}\tilde{F}_0(s)^{1-\Delta} & \text{if } N = 0, \end{cases}$$

$$\mathrm{l}(\infty, p, \tilde{F}_0)(X) \equiv 0.$$

Then the map $(\theta, p, \tilde{F}_0) \mapsto \mathrm{l}(\theta, p, \tilde{F}_0)(X)$ is continuous at every $(\theta, p, \tilde{F}_0)$ such that R_N is a continuity point of $\tilde{F}_0$ or $R_N = 0$. Since the distribution of R_N has at most a point mass at zero, this means at every $(\theta, p, \tilde{F}_0)$ for almost all X.

The log-likelihood is bounded above by

$$\log \sup_{\theta} \theta^N \mathrm{e}^{-\theta R_N} = N \log N - N \log R_N - N$$

(with $0 \log 0 = 0$). This is integrable. Furthermore, the log-likelihood at the true parameters $(\theta_0, p_0, \tilde{F}_{00})$ is bounded below by

$$N \log \theta_0 - \theta_0 T\Delta + \Delta \log p_0 + (1-\Delta)\log(1-p_0) + (1-\Delta)\log\bigl(\mathrm{e}^{-\theta_0 \tau_0}\bigl(1 - \tilde{F}_{00}(R_N)\bigr)\bigr).$$

Here, given $\Delta = 0$, the variable R_N is distributed as $\max(C - E, 0)$ for C and E as in Lemma 1.1 and hence is stochastically bounded above by C. Thus

$$\mathrm{E}\bigl(\log(1 - \tilde{F}_{00}(R_N))|\Delta = 0\bigr) \geqslant \mathrm{E} \log(1 - \tilde{F}_{00}(C)) = \int_0^1 \log u \, \mathrm{d}u > -\infty.$$

Together with the preceding displays, this proves that the supremum of the log-likelihood ratio is integrable above.

The parameters θ and p can be seen to be identifiable from the existence of the consistent estimators $\sum(N_i - 1\{\Delta_i = 0,\ T_i^* > 0\})/\sum T_i^*$ and $n^{-1}\sum \Delta_i$. The parameter $\tilde{F}_0$ is identifiable on its support from the term $(1-\Delta)\int_{[R_N,\infty)} \mathrm{e}^{-\theta s}\, \mathrm{d}\tilde{F}_0(s)$ in the likelihood, since conditionally on $\Delta = 0$ the variable R_N is continuously distributed on its support with a positive density.

Thus we have proved consistency of the maximum likelihood estimator of the parameter $(\theta, p, \tilde{F}_0, \tilde{F}_1)$, relative to the topology introduced previously. Because the distribution function $\tilde{F}_0$ is continuous, the weak topology can be replaced by the uniform topology. This translates into consistency of the estimators for F and G by continuity of the map $(p, \tilde{F}_0, \tilde{F}_1) \mapsto (F, G)$, at least when these distribution functions are restricted to the interval $[0, \tau]$. □

3. Tangent sets and efficient scores

General definitions of tangent spaces (not completely in agreement) are given in Pfanzagl (1982), Bickel *et al*. (1993) and van der Vaart (1998). A tangent space is essentially the (closed) linear span of all score functions of the model. In the present case it is convenient to parametrize the model by $(\theta, p, \tilde{F}_0, \tilde{F}_1)$, where $p = F_1(\infty)$ and $\tilde{F}_0$ and $\tilde{F}_1$ are the probability distributions obtained by renormalizing $\tilde{F}_0$ and $\tilde{F}_1$, and write the density of one observation in the form

$$\mathrm{l}(\theta, p, \tilde{F}_0, \tilde{F}_1)(X) = \theta^{N - 1\{\Delta = 0, T^* > 0\}} \mathrm{e}^{-\theta T^*} \tilde{f}_1(T)^{\Delta} p^{\Delta} (1-p)^{1-\Delta} \tilde{f}_{0,\theta}^*(T^*)^{1-\Delta}.$$

Here $\tilde{f}_{0,\theta}^*(0) = \tilde{F}_{0,\theta}^*\{0\}$ and $\tilde{f}_{0,\theta}^*$ depends on both θ and $\tilde{F}_0$, even though this is not apparent from the notation. We compute score functions for the various parameters separately.

The score function for θ takes the form

$$\dot{l}_{\theta, p, \tilde{F}_0, \tilde{F}_1}(X) = \frac{N - 1\{\Delta = 0,\ T^* > 0\}}{\theta} - T^* + (1-\Delta)\frac{\dfrac{\partial}{\partial \theta} \tilde{f}_{0,\theta}^*(T^*)}{\tilde{f}_{0,\theta}^*(T^*)}.$$

Similarly, the score function for p is given by

$$\dot{\kappa}_{\theta,p,\tilde{F}_0,\tilde{F}_1}(X) = \frac{\Delta}{p} - \frac{1-\Delta}{1-p} = \frac{\Delta - p}{p(1-p)}.$$

The distribution $\tilde{F}_1$ is completely unknown and is included in the likelihood through the multiplicative factor $\tilde{f}_1(T)^{\Delta}$. For a bounded, measurable function a such that $\tilde{F}_1 a = 0$, the definition $\mathrm{d}\tilde{F}_{1s}(t) = (1 + sa(t))\mathrm{d}\tilde{F}_1(t)$ defines a probability distribution for every s that is sufficiently close to 0. Inserting this 'path' in the likelihood and differentiating at $s = 0$ yields the score function

$$A_{\theta,p,\tilde{F}_0,\tilde{F}_1} a(X) = \Delta a(T).$$

This set of functions forms a linear space if a ranges over a linear space. Since the operator $A_{\theta,p,\tilde{F}_0,\tilde{F}_1} : L_2(\tilde{F}_1) \mapsto L_2(P_{\theta,p,\tilde{F}_0,\tilde{F}_1})$ is continuous, the closed linear span of all score functions will contain all measurable functions of this type such that $\tilde{F}_1 a = 0$ and $\tilde{F}_1 a^2 < \infty$.

Computing score functions for $\tilde{F}_0$ is more involved. This parameter is hidden in the distribution $\tilde{F}^*_{0,\theta}$, which is characterized in Lemma 1.1. Suppose that $b\colon [0, \infty) \mapsto \mathbb{R}$ is a bounded, non-increasing function that is caglad on $(0, \infty)$ such that $\tilde{F}^*_{0,\theta} b = 0$, and set

$$g_s(t) = \tilde{f}^*_{0,\theta}(t)(1 + sb(t)),$$

where we interpret $g_s(0)$ as a point mass at 0. Then g_s is non-negative for sufficiently small $|s|$, and for $s \geqslant 0$ the function $t \mapsto \mathrm{e}^{-\theta t} g_s(t)$ is non-increasing on $(0, \infty)$ and satisfies $\theta g_s(0) \geqslant g_s(0+)$. Therefore, by Lemma 1.1, G_s is a distribution that can be written in the form of a distribution $\tilde{F}^*_{0,\theta}$ for some $\tilde{F}_0$ (for every fixed θ). If we insert this distribution in the log-likelihood and differentiate (from the right) with respect to s at $s = 0$, then we find the score function

$$B_{\theta,p,\tilde{F}_0,\tilde{F}_1} b(X) = (1 - \Delta) b(T^*).$$

Even though this may not be a score function for a two-sided ('regular') submodel, we consider this function as a member of the tangent set (thus deviating from the basic definition of Bickel *et al.* 1993). The linear span of the tangent set will contain all functions $(1 - \Delta)b(T^*)$ such that $F^*_{0,\theta} b = 0$ and such that b is bounded, caglad and of bounded variation. Since the operator $B_{\theta,p,\tilde{F}_0,\tilde{F}_1} : L_2(\tilde{F}^*_{0,\theta}) \mapsto L_2(P_{\theta,p,\tilde{F}_0,\tilde{F}_1})$ is continuous, the closure of the linear span contains all measurable functions $(1 - \Delta)b(T^*)$ such that $\tilde{F}^*_{0,\theta} b = 0$ and $\tilde{F}^*_{0,\theta} b^2 < \infty$. These observations readily yield the following lemma.

Lemma 3.1. *The efficient score function for θ is given by*

$$\tilde{l}_{\theta,p,\tilde{F}_0,\tilde{F}_1}(X) = \frac{N - 1\{\Delta = 0,\ T^* > 0\}}{\theta} - T^*.$$

Proof. For the function $\tilde{l}_{\theta,p,\tilde{F}_0,\tilde{F}_1}$ as defined by the preceding display, the difference $\dot{l}_{\theta,p,\tilde{F}_0,\tilde{F}_1} - \tilde{l}_{\theta,p,\tilde{F}_0,\tilde{F}_1}$ has the form $(1 - \Delta)b(T^*)$ for some function b and hence is a score function for $\tilde{F}_0$, provided that it is square-integrable. The latter can be verified. It suffices to show that $\tilde{l}_{\theta,p,\tilde{F}_0,\tilde{F}_1}$ is orthogonal to the set of scores for p, $\tilde{F}_0$ and $\tilde{F}_1$.

Given T^* and $\Delta = 1$, the variable N is Poisson distributed with mean θT^*. Furthermore, given T^*, $T^* > 0$ and $\Delta = 0$, the variable $N - 1$ is Poisson distributed with mean θT^*. It follows that

$$\mathrm{E}\left(\frac{N - 1\{\Delta = 0,\ T^* > 0\}}{\theta} - T^* | T^*,\ \Delta\right) = \frac{\theta T^*}{\theta} - T^* = 0.$$

This proves that $\tilde{l}_{\theta, p, \tilde{F}_0, \tilde{F}_1}$ is orthogonal to all functions of (T^*, Δ). □

It is interesting that the efficient score function for θ does not depend on the nuisance parameters p, $\tilde{F}_0$, $\tilde{F}_1$. This unusual situation makes it possible to use the 'efficient score equation'

$$\sum_{i=1}^{n} \tilde{l}_{\theta, p, \tilde{F}_0, \tilde{F}_1}(X_i) = 0$$

to define an estimator for θ. This yields the conditional likelihood estimator given by (1.6). The fact that this estimator solves the efficient score equation suggests that it is asymptotically efficient in the semi-parametric sense. This is indeed the case, as can be proved by analysing its asymptotic properties directly with the help of the delta method, or by general arguments based on linearizing the efficient score equation (cf. van der Vaart 1996; or 1998, Section 25.8).

4. Asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta)$

In this section we prove the asymptotic normality of the maximum likelihood estimator $\hat{\theta}_n$ for θ by showing that it is asymptotically equivalent to the *ad hoc* estimator (1.6). We assume that there exists an interval $[0, \tau_0]$ on which F has a positive, continuous density and $G(\tau_0) < 1$ and such that the measure with density $(1 - G)\mathrm{d}F$ gives zero mass to (τ_0, ∞). One possible case of interest in which this is true is when F has a continuous, positive density on a support $[0, \tau_0]$ and $G(\tau_0) < 1$, indicating that a positive fraction of people is not censored (i.e. does not move).

Theorem 4.1. *Suppose that the true values of F and G satisfy the stated conditions. Then $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_{n,c}) \xrightarrow{\mathrm{P}} 0$. Consequently, the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with mean zero and variance the inverse of the efficient Fisher information for θ.*

Proof. The maximum likelihood estimator maximizes the profile log-likelihood function, obtained by maximizing the log-likelihood over all parameters (F_0, F_1), for fixed θ. The term involving F_1 does not depend on θ and hence can be dropped from the profile likelihood. By Lemma 1.2 and expression (1.5) for the likelihood, the remaining part of the profile log-likelihood can be written in the form

$$\sum_{i=1}^{n}((N_i - 1\{\Delta_i = 0,\ T_i^* > 0\})\log\theta - \theta T_i^*) + \sum_{i=1}^{n}(1-\Delta_i)\log(\theta^{1\{T_i^*>0\}}\mathrm{e}^{\theta T_i^*}\hat{h}_{0,\theta}(\mathrm{e}^{\theta T_i^*})),$$

where $\hat{H}_{0,\theta}$ is the left derivative of the least concave majorant of the function $\mathbb{H}_{0,\theta}$ given in Lemma 1.2. Since $\hat{\theta}$ maximizes this expression, it is a zero of its derivative, if this exists at $\hat{\theta}$. Its existence is proved in Lemma 4.2. Setting the derivative of the first term to zero yields the efficient score equation, which is solved by $\hat{\theta}_{n,c}$ given by (1.6). We shall show that the derivative of the second term is asymptotically negligible. If we denote the derivative of this term by $D_n(\theta)$, then it follows that

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(N_i - 1\{\Delta_i = 0,\ T_i^* > 0\})}{\sum_{i=1}^{n} T_i^* - D_n(\hat{\theta})}. \tag{4.1}$$

The theorem is proved once it is shown that $D_n(\hat{\theta}) = o_P(\sqrt{n})$.

To simplify notation, we write the formulae as if $\Delta_1 = \cdots = \Delta_n = 0$ and $T_1^* \leqslant T_2^* \leqslant \ldots \leqslant T_n^*$. Since the subdistribution function $H_{0,\theta}$ is continuous on $(1, \infty)$ with an atom at 1, the values $\mathrm{e}^{\theta T_i^*}$ will almost surely be tied at 1 only. The least concave majorant $\hat{H}_{0,\theta}$ of $\mathbb{H}_{0,\theta}$ is piecewise linear and changes direction only at points where $\hat{H}_{0,\theta}$ and $\mathbb{H}_{0,\theta}$ are equal. The point 1 may or may not be one of these points. Define $A_\theta = \{i\colon \mathbb{H}_{0,\theta}(\mathrm{e}^{\theta T_i^*}) = \hat{H}_{0,\theta}(\mathrm{e}^{\theta T_i^*}),\ T_{i+1}^* > 0\}$ as the set of indices of the points $\mathrm{e}^{\theta T_i^*}$ where $\hat{H}_{0,\theta}$ and $\mathbb{H}_{0,\theta}$ coincide, where in the case $\hat{H}_{0,\theta}(1) = \mathbb{H}_{0,\theta}(1)$ only the largest index of the T_i^* that are tied at zero is included. Furthermore, define $k_{i,\theta} = \min\{j \in A_\theta\colon j > i\}$ for $i < n$; this is the index of the smallest point $\mathrm{e}^{\theta T_j^*}$ larger than $\mathrm{e}^{\theta T_i^*}$ where $\hat{H}_{0,\theta}$ and $\mathbb{H}_{0,\theta}$ coincide. (This minimum always exists, since $n \in A_\theta$.) For $i = 0$, the number $k_{0,\theta} = \min\{j\colon j \in A_\theta\}$ is the index $j \in A_\theta$ of the smallest point $\mathrm{e}^{\theta T_j^*}$ where $\hat{H}_{0,\theta}$ and $\mathbb{H}_{0,\theta}$ coincide. (This point may be 1 or bigger than 1; in the first case it is the largest index of the points tied at 1.) The slope of the least concave majorant at $\mathrm{e}^{\theta T_j^*}$ is constant for $i < j \leqslant k_{i,\theta}$ for every $i \in A_\theta$ and easily computed as the quotient of the increase in $\mathbb{H}_{0,\theta}$ over the interval $(\exp(\theta T_i^*), \exp(\theta T_{k_{i,\theta}}^*)]$ and the length of this interval. The second term of the profile likelihood can be rewritten as

$$\#\{i\colon T_i^* > 0\}\log\theta + \sum_{j=1}^{k_{0,\theta}} \log\left(\frac{k_{0,\theta}}{n}\frac{\mathrm{e}^{\theta T_j^*}}{\mathrm{e}^{\theta T_{k_{0,\theta}}^*}}\right) + \sum_{\substack{i\in A_\theta\\ i<n}}\sum_{j=i+1}^{k_{i,\theta}} \log\left(\frac{k_{i,\theta}-i}{n}\frac{\mathrm{e}^{\theta T_j^*}}{\mathrm{e}^{\theta T_{k_{i,\theta}}^*} - \mathrm{e}^{\theta T_i^*}}\right).$$

If we perturb $\hat{\theta}$ slightly then the points $\mathrm{e}^{\hat{\theta}T_i^*}$ change location slightly and the graph of the empirical subdistribution function $\mathbb{H}_{0,\hat{\theta}}$ is deformed slightly as well. The graph of the concave majorant $\hat{H}_{0,\hat{\theta}}$ is obtained by linearly connecting the points $(\mathrm{e}^{\hat{\theta}T_i^*}, \mathbb{H}_{0,\hat{\theta}}(\mathrm{e}^{\hat{\theta}T_i^*}))$ for $i \in A_{\hat{\theta}}$. If we perturb $\hat{\theta}$ slightly into θ', then $\mathbb{H}_{0,\theta'}(\mathrm{e}^{\theta' T_i^*}) = \mathbb{H}_{0,\hat{\theta}}(\mathrm{e}^{\hat{\theta}T_i^*})$ for every i, because the points $\mathrm{e}^{\theta T_i^*}$ do not change order if θ moves from $\hat{\theta}$ to θ'. Suppose that the slope of $\hat{H}_{0,\hat{\theta}}$ decreases strictly at every point $\mathrm{e}^{\hat{\theta}T_i^*}$ at which $\hat{H}_{0,\hat{\theta}}$ and $\mathbb{H}_{0,\hat{\theta}}$ coincide (i.e. for every $i \in A_{\hat{\theta}}$). Then the graph obtained by connecting the points $(\mathrm{e}^{\theta' T_i^*}, \mathbb{H}_{0,\hat{\theta}}(\mathrm{e}^{\hat{\theta}T_i^*}))$ linearly is concave and a majorant of $\mathbb{H}_{0,\theta'}$, provided θ' is sufficiently close to $\hat{\theta}$, and hence is equal to the least concave

majorant $\hat{H}_{0,\theta'}$. This shows that in this case the set $A_{\theta'}$ coincides with $A_{\hat{\theta}}$. The other case is that the slope of the concave majorant $\hat{H}_{0,\hat{\theta}}$ does not decrease strictly at every point $e^{\hat{\theta}T_i^*}$ at which $\hat{H}_{0,\hat{\theta}}$ and $\mathbb{H}_{0,\hat{\theta}}$ coincide. Then three or more points $e^{\hat{\theta}T_i^*}$ are on a straight line. If we perturb $\hat{\theta}$ slightly and make it smaller, then A_θ may change, but if we make $\hat{\theta}$ bigger, then A_θ does not change when θ moves from $\hat{\theta}$ to θ'. By Lemma 4.2 below, the profile likelihood is differentiable at $\hat{\theta}$. In view of the preceding observations, if we compute its derivative from the right at $\hat{\theta}$, then we may set A_θ in the preceding display equal to $\hat{A} := A_{\hat{\theta}}$ and hence set $k_{i,\theta}$ equal to $\hat{k}_i := k_{i,\hat{\theta}}$. This derivative is equal to

$$\frac{\#\{i\colon T_i^* > 0\}}{\hat{\theta}} + \sum_{j=1}^{\hat{k}_0}(T_j^* - T_{\hat{k}_0}^*) - \sum_{\substack{i\in\hat{A}\\ i<n}}\sum_{j=i+1}^{\hat{k}_i}\frac{(T_{\hat{k}_i}^* - T_j^*)e^{\hat{\theta}(T_{\hat{k}_i}^* - T_j^*)} - (T_i^* - T_j^*)e^{\hat{\theta}(T_i^* - T_j^*)}}{e^{\hat{\theta}(T_{\hat{k}_i}^* - T_j^*)} - e^{\hat{\theta}(T_i^* - T_j^*)}}.$$

Using Taylor type arguments we can see that there exist a neighbourhood of 0 and a constant C such that, for every u and v in the neighbourhood,

$$\left|\frac{ve^v - ue^u}{e^v - e^u} - 1 - \frac{u+v}{2}\right| \leqslant C(u^2 + v^2).$$

We use this with $v = \hat{\theta}(T_{\hat{k}_i}^* - T_i^*)$ and $u = \hat{\theta}(T_i^* - T_j^*)$ to expand the third term. By Corollary 8.3 and the fact that the logarithm is Lipschitz on $[1, \infty)$, $\max_i|T_{\hat{k}_i}^* - T_i^*| = o_P(n^{-1/4})$. Since there are at most $n - \hat{k}_0 \leqslant n$ terms in the double sum, the derivative is up to a term of order $o_P(\sqrt{n})$ asymptotically equivalent to

$$\begin{aligned}\#\{i\colon T_i^* > 0\}\frac{1}{\hat{\theta}} + \sum_{j=1}^{\hat{k}_0}(T_j^* - T_{\hat{k}_0}^*) - \sum_{\substack{i\in\hat{A}\\ i<n}}\sum_{j=i+1}^{\hat{k}_i}\left(\frac{1}{\hat{\theta}} + \tfrac{1}{2}(T_{\hat{k}_i}^* + T_i^*) - T_j^*\right)\\ = \frac{1}{\hat{\theta}}(\hat{k}_0 - \hat{z} - \hat{z}T_{\hat{k}_0}^*\hat{\theta}) + \sum_{j=\hat{z}+1}^{\hat{k}_0}(T_j^* - T_{\hat{k}_0}^*) - \sum_{\substack{i\in\hat{A}\\ i<n}}\sum_{j=i+1}^{\hat{k}_i}\left(\tfrac{1}{2}(T_{\hat{k}_i}^* + T_i^*) - T_j^*\right),\end{aligned} \tag{4.2}$$

where $\hat{z}$ is the number of T_i^* tied at 0. We can conclude the proof by showing that this expression is $o_P(\sqrt{n})$. By Corollaries 8.3 and 8.4 the sum in the second term has of the order of $O_P(n^{2/3}(\log n)^{1/3})$ terms of maximal order $O_P(n^{-1/3}(\log n)^{1/3})$, and hence it is certainly of order $o_P(\sqrt{n})$. We consider the first and third terms separately.

Under some conditions, the third term of the right-hand side of (4.2) can be shown to be $o_P(\sqrt{n})$ as a consequence of Corollary 8.3 without using the exact definitions of the $\hat{k}_i$, but we shall give a proof based on Lemma 7.2. In any case the cancellation of positive and negative terms in the sum is essential. Let $S_i = e^{\hat{\theta}T_i^*}$. Then it suffices to prove the analogous property for the S_i instead of the T_i^*, because the difference between $\frac{1}{2}(S_i + S_{\hat{k}_i}) - S_j$ and its linearization in $\frac{1}{2}(T_i^* + T_{\hat{k}_i}^*) - T_j^*$ is of order $O_P(n^{-2/3}(\log n)^{2/3})$, uniformly in i and j, by Corollary 8.3. Because $\hat{H}_{0,\theta}$ is linear between S_i and $S_{\hat{k}_i}$ with slope $((\hat{k}_i - i)/n)/(S_{\hat{k}_i} - S_i)$, we have

$$\sum_{j=i+1}^{\hat{k}_i}(\tfrac{1}{2}(S_i+S_{\hat{k}_i})-S_j) = n(S_{\hat{k}_i}-S_i)\frac{1}{\hat{k}_i-i}\sum_{j=i+1}^{\hat{k}_i}(\hat{H}_{0,\hat{\theta}}(\tfrac{1}{2}(S_i+S_{\hat{k}_i}))-\hat{H}_{0,\hat{\theta}}(S_j))$$

$$= n(S_{\hat{k}_i}-S_i)\left[\frac{1}{\hat{k}_i-i}\sum_{j=i+1}^{\hat{k}_i}(\mathbb{H}_{0,\hat{\theta}}(\tfrac{1}{2}(S_i+S_{\hat{k}_i}))-\mathbb{H}_{0,\hat{\theta}}(S_j))+o_P\left(\frac{1}{\sqrt{n}}\right)\right],$$

uniformly in i, by Lemma 7.2. The average in the square brackets can be computed explicitly and, with l_i and u_i being the number of indices $i<j\leqslant\hat{k}_i$ such that $\frac{1}{2}(S_i+S_{\hat{k}_i})>S_j$ or $\frac{1}{2}(S_i+S_{\hat{k}_i})<S_j$, respectively, is equal to

$$\frac{1}{2n}(l_i-u_i)\frac{\hat{k}_i-i+1}{\hat{k}_i-i} = (\mathbb{H}_{0,\hat{\theta}}-\hat{H}_{0,\hat{\theta}})(\tfrac{1}{2}(S_i+S_{\hat{k}_i}))\frac{\hat{k}_i-i+1}{\hat{k}_i-i},$$

where the equality follows from drawing a picture of $\mathbb{H}_{0,\hat{\theta}}$ and $\hat{H}_{0,\hat{\theta}}$. This is $o_P(n^{-1/2})$ uniformly in i, by Lemma 7.2. We conclude that

$$\left|\sum_{\substack{i\in\hat{A}\\ i<n}}\sum_{j=i+1}^{\hat{k}_i}(\tfrac{1}{2}(S_i+S_{\hat{k}_i})-S_j)\right| \leqslant \sum_{\substack{i\in\hat{A}\\ i<n}}\sqrt{n}(S_{\hat{k}_i}-S_i)o_P(1).$$

This is of the desired order, because the sum telescopes out to the (finite) length of the support.

Finally, we consider the first term on the right-hand side of (4.2). By the definition of the concave majorant and Taylor's theorem, we have

$$n\mathrm{e}^{\hat{\theta}T^*_{\hat{k}_0}}(\hat{H}_{0,\hat{\theta}}(1)-\mathbb{H}_{0,\hat{\theta}}(1)) = n\left(\frac{\hat{k}_0}{n}-\frac{\hat{z}}{n}\mathrm{e}^{\hat{\theta}T^*_{\hat{k}_0}}\right) = \hat{k}_0-\hat{z}-\hat{z}\hat{\theta}T^*_{\hat{k}_0}+o_P(\sqrt{n}),$$

since $(T^*_{\hat{k}_0})^2$ is of order $o_P(n^{-1/2})$ by Corollary 8.3. If the left-hand side of this display is of order $o_P(\sqrt{n})$, then it follows that the first term of the right of (4.2) is of order $o_P(\sqrt{n})$ as well, and the proof is complete. The left-hand side is of order $o_P(\sqrt{n})$ if the second assertion of Lemma 7.2 is true. To be able to apply this assertion we must first show that $\sqrt{n}(\hat{\theta}-\theta)\geqslant O_P(1)$. By the definition of $\hat{H}_{0,\hat{\theta}}$ the left-hand side of the preceding display is certainly non-negative, whence we have proved that $D_n(\hat{\theta})\geqslant -o_P(\sqrt{n})$ (where the minus is superfluous but aids the interpretation). In view of (4.1), it follows that $\hat{\theta}\geqslant\hat{\theta}_{n,c}+o_P(n^{-1/2})\geqslant\theta+O_P(n^{-1/2})$. The proof is complete. □

Lemma 4.2. *The profile likelihood function is differentiable at the maximum likelihood estimator* $\hat{\theta}$.

Proof. The profile log-likelihood function is proportional to the function

$$\theta\mapsto\sum_{i=1}^{n}(N_i\log\theta-\Delta_iT^*_i\theta)+\sum_{i=1}^{n}((1-\Delta_i)\log\hat{h}_{0,\theta}(\mathrm{e}^{\theta T^*_i})).$$

The first part is continuously differentiable. The second part is both left- and right-

differentiable and is strictly convex. In particular, its left derivative is strictly smaller than its right derivative whenever the two derivatives are not equal.

Let f_1 and f_2 be the two parts of the profile log-likelihood function. Since the maximum likelihood estimator $\hat{\theta}$ maximizes the profile log-likelihood, we have

$$f_1'(\hat{\theta}-) + f_2'(\hat{\theta}-) \geqslant 0 \geqslant f_1'(\hat{\theta}+) + f_2'(\hat{\theta}+).$$

By continuity of f_1' it follows that $f_2'(\hat{\theta}-) \geqslant f_2'(\hat{\theta}+)$. Combined with the preceding paragraph, this implies that $f_2'(\hat{\theta}-) = f_2'(\hat{\theta}+)$. □

5. Estimation of Λ_F and F

In this section we prove the asymptotic normality of the maximum likelihood estimators of Λ_F and F. The arguments are much like those in Gill (1997), except that we must deal with an additional parameter θ. A realization of the maximum likelihood estimator of F can be seen in Figure 2.

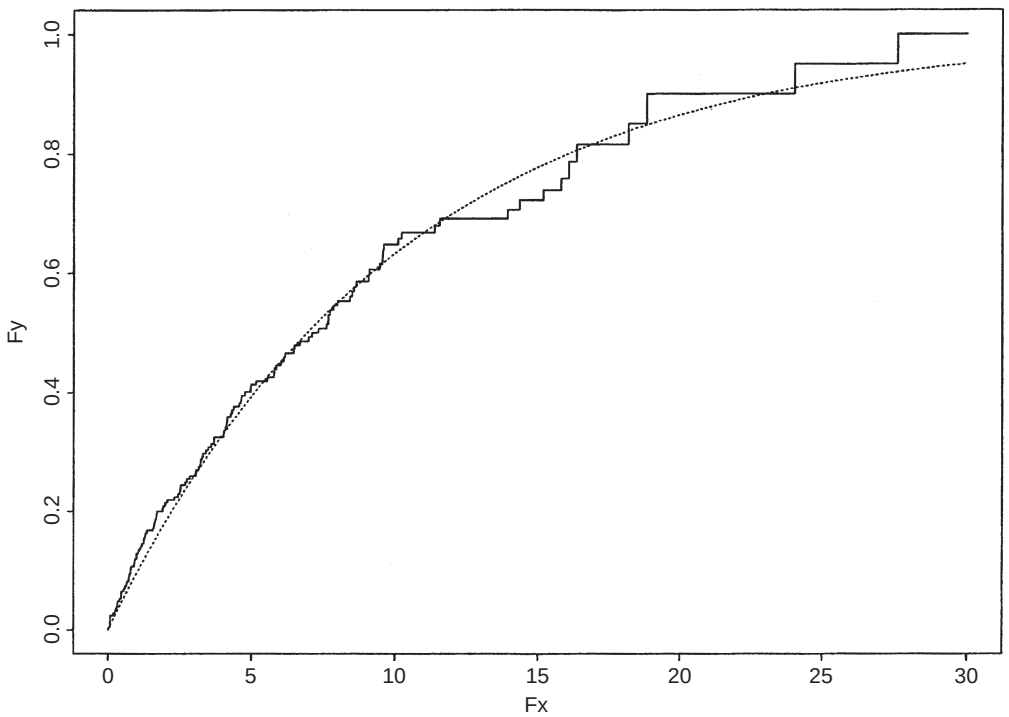

Figure 2. A realization of the maximum likelihood estimator of F based on 250 observations (step function) and the true distribution function, the exponential distribution with intensity 0.1 (dotted line).

In analogy with the notation $H_{0,\theta}$, let $H_{1,\theta}(t) = P_{\theta,F,G}(e^{\theta T^*} \leqslant t, \Delta = 1)$. The maximum likelihood estimator of this parameter, for fixed θ, is the empirical subdistribution function $\mathbb{H}_{1,\theta}$ of the variables $e^{\theta T_i}$ with $\Delta_i = 1$.

The cumulative hazard function Λ_F can be expressed in the distributions $H_{0,\theta}$ and $H_{1,\theta}$ as

$$\Lambda_F(t) = \int_{[1,e^{\theta t}]} \frac{\mathrm{d}H_{1,\theta}}{1 - H_{1,\theta-} - H_{0,\theta-} + \mathrm{id}h_{0,\theta-}},$$

where $\mathrm{id}(y)$ is the identify function $y \mapsto y$ and $h_{0,\theta-} = h_{0,\theta}$ on $(1, \infty)$, but $h_{0,\theta}(1-)$ is defined to be 0. (Together with our earlier conventions, this means that the values $h_{0,\theta}(1-)$, $h_{0,\theta}(1) = H_{0,\theta}\{1\}$ and $h_{0,\theta}(1+)$ may all be different for a general parameter in the model, even though we shall assume that $h_{0,\theta}(1+) = h_{0,\theta}(1)$ for the true value of $h_{0,\theta}$.) The maximum likelihood estimator $\hat{\Lambda}_F$ is obtained by replacing the parameter (θ, $H_{0,\theta}$, $H_{1,\theta}$, $h_{0,\theta}$) by its maximum likelihood estimator ($\hat{\theta}$, $\hat{H}_{0,\hat{\theta}}$, $\hat{H}_{1,\hat{\theta}}$, $\hat{h}_{0,\hat{\theta}}$). We consider estimating Λ_F on an interval $[0, \tau]$ such that $((1 - F)(1 - G))(\tau) > 0$.

Theorem 5.1. *Let the conditions of Theorem 4.1 hold. Then the sequences of processes* $\sqrt{n}(\hat{\Lambda}_F - \Lambda_F)$ *and* $\sqrt{n}(\hat{F} - F)$ *converge in distribution to tight Gaussian limits in* $l^\infty[0, \tau]$ *for every* $\tau < \tau_0$.

We obtain the asymptotic distribution of $\hat{\Lambda}_F$ and $\hat{F}$ essentially by the delta method, but need to work hard to handle the term $\hat{h}_{0,\hat{\theta}}$. Unlike the estimators $\hat{\theta}$, $\hat{H}_{0,\hat{\theta}}$, $\hat{H}_{1,\hat{\theta}}$, this estimator does not converge at rate $\sqrt{n}$, but at rate $n^{1/3}$ and only in a pointwise sense. For that reason the $\sqrt{n}$ rates of $\hat{\Lambda}_F$ and $\hat{F}$ are not at all obvious, and necessitate a long proof. This proof is contained in Section 9.

6. Estimation of Λ_G and G

The cumulative hazard function Λ_G can be expressed in $H_{0,\theta}$ and $H_{1,\theta}$ as

$$\Lambda_G(t) = -\int_{[1,e^{\theta t}]} \frac{\mathrm{id}\,\mathrm{d}h_{0,\theta}}{1 - H_{0,\theta-} - H_{1,\theta} + \mathrm{id}h_{0,\theta-}}.$$

Here $-\mathrm{d}\,h_{0,\theta}(1)$ is defined as $H_{0,\theta}(1) - h_{0,\theta}(1+) = h_{0,\theta}(1) - h_{0,\theta}(1+)$, the downward jump of $h_{0,\theta}$ at 1, and $h_{0,\theta}(1-) = 0$. The maximum likelihood estimator $\hat{\Lambda}_G$ is obtained by replacing (θ, $H_{0,\theta}$, $H_{1,\theta}$) by its maximum likelihood estimator. Unlike $\hat{\Lambda}_F$, the estimator $\hat{\Lambda}_G$ is dominated by the estimator $\hat{h}_{0,\hat{\theta}}$ and its rate of convergence is slower than $n^{-1/2}$. The following theorem gives the rate for the uniform norm.

Theorem 6.1. *Let the conditions of Theorem 4.1 hold and suppose that F is twice differentiable with bounded second derivative and that G has a bounded density on* $[0, \tau_0]$. *The sequences* $\|\hat{\Lambda}_G - \Lambda_G\|_\infty$ *and* $\|\hat{G} - G\|_\infty$ *are* $O_P((n/\log n)^{-1/3})$, *if* $\|\cdot\|_\infty$ *is the uniform norm on the interval* $[0, \tau]$, *for every* $\tau < \tau_0$.

The discussion ahead suggests that this uniform rate of convergence is sharp, but also that the sequence $\hat{\Lambda}_G(t) - \Lambda_G(t)$ converges at rate $O_P(n^{-1/3})$ for every fixed t. Our derivation of the latter result is incomplete. Furthermore, even if the sequence $n^{1/3}(\hat{\Lambda}_G - \Lambda_G)(t)$ converges in distribution, it is not clear that the sequence $(\hat{G} - G)(t)$ also converges at the rate $O_P(n^{-1/3})$, because the transition from hazard function to survival function appears to require some uniformity if the functions involved possess jumps. Claims to this effect, made in passing by Gill (1997), appear to be without proof at this time.

The problems encountered here appear to be caused by the jumps in the maximum likelihood estimator $\hat{h}_{0,\hat{\theta}}$ and may be real. We can remedy this by using a smoothed version of $\hat{h}_{0,\hat{\theta}}$ instead. For continuous approximations $\tilde{h}_{0,\theta}$ to $\hat{h}_{0,\theta}$, consider the estimators

$$\tilde{\Lambda}_G(t) = -\int_{[1,e^{\hat{\theta}t}]} \frac{\mathrm{id}\, d\tilde{h}_{0,\hat{\theta}}}{1 - \hat{H}_{0,\hat{\theta}-} - \hat{H}_{1,\hat{\theta}} + \mathrm{id}\, \tilde{h}_{0,\hat{\theta}-}}.$$

We can construct the modifications $\tilde{h}_{0,\theta}$ in many ways, the simplest perhaps being kernel smoothing. Here we send the bandwidth to zero faster than $n^{-1/3}$ and use a special kernel to ensure that $\tilde{h}_{0,\theta}$ is supported on $[1, \infty)$, as is $\hat{H}_{0,\theta}$. For instance,

$$\tilde{h}_{0,\theta}(t) = \begin{cases} \hat{h}_{0,\theta} * U[-a_n, a_n](t), & t > 1 + a_n, \\ \hat{h}_{0,\theta} * U[1 - t, a_n](t), & 1 \leqslant t \leqslant 1 + a_n, \end{cases}$$

where $U[a, b]$ is the uniform measure on $[a, b]$. We let $\tilde{H}_{0,\theta}(t) = \hat{H}_{0,\theta}(1) + \int_0^t \tilde{h}_{0,\theta}(s)\, \mathrm{d}s$.

Theorem 6.2. *Let the conditions of Theorem 6.1 hold and suppose that $a_n = o(n^{-1/3})$. Then, for every fixed $t < \tau_0$, the sequence $n^{1/3}(\tilde{\Lambda}_G - \Lambda_G)(t)$ converges in distribution to*

$$\frac{-4^{1/3} e^{\theta t} g^{1/3}(t) \int_{[t,\infty)} e^{-s\theta}(1 - F)(s) \mathrm{d}G(s)^{1/3}}{(1 - F)^{2/3}(t)(1 - G)(t)} \operatorname*{argmax}_{h \in \mathbb{R}} \{ Z(h) - h^2 \},$$

for Z a standard Brownian motion. Consequently, the sequence $n^{1/3}(\tilde{G} - G)(t)$ converges in distribution as well.

The proofs of Theorems 6.1 and 6.2 are contained in Section 10. A realization of the maximum likelihood estimator of G can be seen in Figure 3.

7. Asymptotics for $\hat{H}_{0,\hat{\theta}}$ and $\hat{H}_{1,\hat{\theta}}$.

In this section we prove the asymptotic normality of the maximum likelihood estimators of the subdistribution functions $H_{0,\theta}$ and $H_{1,\theta}$. Let $\|\cdot\|_{[a,b]}$ denote the supremum norm on $[a, b]$.

Lemma 7.1. *The sequence of processes $\sqrt{n}(\mathbb{H}_{0,\tilde{\theta}}(t) - H_{0,\theta}(t^{\theta/\tilde{\theta}}),\ \mathbb{H}_{1,\tilde{\theta}}(t) - H_{1,\theta}(t^{\theta/\tilde{\theta}}))$ converges for any sequence $\tilde{\theta} \xrightarrow{\mathrm{P}} \theta$ in distribution in $l^\infty[1, \sigma] \times l^\infty[1, \sigma]$ to a Gaussian process with continuous sample paths.*

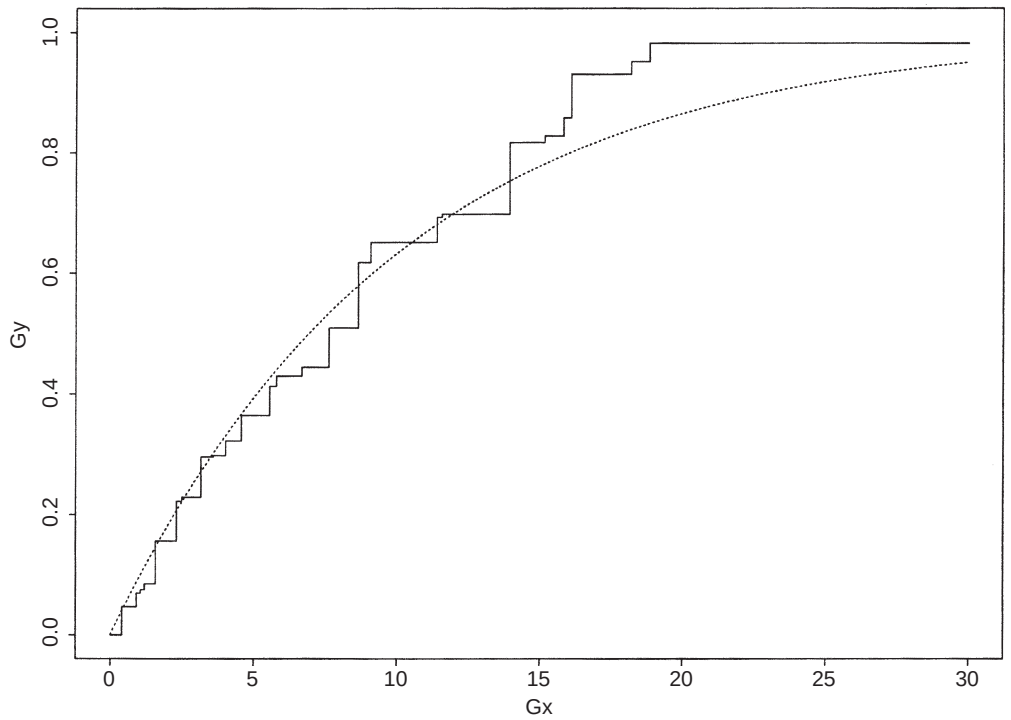

Figure 3. A realization of the maximum likelihood estimator of G based on 250 observations (step function) and the true distribution function, the exponential distribution with rate 0.1 (dotted line).

Proof. The subdistribution functions $\mathbb{H}_{\delta,\tilde{\theta}}(t)$ can be written as $\mathbb{F}^*_\delta(\log t/\tilde{\theta})$ for the subdistribution functions $\mathbb{F}^*_\delta(t) = n^{-1}\sum_{i=1}^n 1\{T_i^* \leqslant t, \Delta_i = \delta\}$. The empirical processes $\sqrt{n}(\mathbb{F}^*_\delta - F^*_{\delta,\theta})$ converge in the space $l^\infty[0, \infty]$ to tight Gaussian processes with uniformly continuous sample paths. The lemma now follows by Lemma 9.1. □

In the following lemma we show the asymptotic equivalence of the empirical distribution and its least concave majorant. For the ordinary empirical distribution this is a well-known property of the Grenander estimator. See, for example, Robertson *et al*. (1988). Gill (1997) extends this to the least concave majorant of an arbitrary discrete estimator of a concave distribution function and also considers the special situation of distribution functions starting with a point mass at 1 needed in this paper. We further extend his result to randomly placed 'observations' $e^{\tilde{\theta}T_i^*}$.

Lemma 7.2. *If $H_{0,\theta}$ is continuous and strictly concave on its support $[1, \sigma_0]$ and $\tilde{\theta} \xrightarrow{P} \theta$, then $\sqrt{n}\|\mathbb{H}_{0,\tilde{\theta}} - \hat{H}_{0,\tilde{\theta}}\|_{[\hat{\sigma}_1,\sigma]} \xrightarrow{P} 0$ for every $\sigma > 1$, where $\hat{\sigma}_1$ is the smallest $t \geqslant 1$ where $t \mapsto \hat{H}_{0,\tilde{\theta}}(t)$ changes direction. If $\sqrt{n}(\tilde{\theta} - \theta) \geqslant O_P(1)$ and $H_{0,\theta}$ is continuously differentiable, then we also have that $\sqrt{n}\|\mathbb{H}_{0,\tilde{\theta}} - \hat{H}_{0,\tilde{\theta}}\|_{[1,\sigma]} \xrightarrow{P} 0$ for every $\sigma > 1$.*

Proof. Let $\hat{\underline{H}}_{0,\tilde{\theta}}$ be the smallest function that is zero on [0, 1], strictly positive and concave on $(1, \infty)$ and majorizes the function $\underline{\mathbb{H}}_{0,\tilde{\theta}} := \mathbb{H}_{0,\tilde{\theta}} - \mathbb{H}_{0,\tilde{\theta}}(1)$. Then $\hat{\underline{H}}_{0,\tilde{\theta}} = \hat{H}_{0,\tilde{\theta}} - \mathbb{H}_{0,\tilde{\theta}}(1)$ on $[\hat{\sigma}_1, \infty)$, and hence it suffices to show that $\sqrt{n}\|\hat{\underline{H}}_{0,\tilde{\theta}} - \underline{\mathbb{H}}_{0,\tilde{\theta}}\|_{[\hat{\sigma}_1,\sigma]} \to 0$ in probability for every $\sigma > 1$.

We can do this by adapting the proof of Theorem 1 in the Appendix of Gill (1997). An inessential difference from the situation covered by Gill's Theorem 1 is that our empirical distribution $\underline{\mathbb{H}}_{0,\tilde{\theta}}$ and concave majorant start at 1. This can be accommodated by simply shifting the axis. A more important difference is that we shall centre $\underline{\mathbb{H}}_{0,\tilde{\theta}}(t)$ (i.e. Gill's F_n) at the random distribution function $\underline{H}_{\tilde{\theta}}(t) := H_{0,\theta}(t^{\theta/\tilde{\theta}}) - \mathbb{H}_{0,\theta}(1)$ rather than at a fixed distribution (as Gill's F). With this centring we do have that $\sqrt{n}(\underline{\mathbb{H}}_{0,\tilde{\theta}} - \underline{H}_{\tilde{\theta}})$ converges in distribution on compacta in $[1, \infty]$ to a Gaussian process with continuous sample paths, whenever $\tilde{\theta} \xrightarrow{P} \theta$ (i.e. Gill's assumption of convergence of $\sqrt{n}(F_n - F)$). If we now pass to an almost sure representation of the weakly converging sequence $(\sqrt{n}(\underline{\mathbb{H}}_{0,\tilde{\theta}} - \underline{H}_{\tilde{\theta}}), \tilde{\theta})$ in $l^\infty[1, \sigma] \times \mathbb{R}$, then we are almost back in Gill's situation, except for the fact that Gill's F becomes a sequence depending on n in our situation. It can now be seen that Gill's proof remains valid and gives the desired result, where we use the fact that $\underline{H}_{\tilde{\theta}} \to \underline{H}_{\theta}$ uniformly on compacta, whence $\underline{H}_{\tilde{\theta}}$ will be a (strictly) concave function eventually.

To prove the second assertion of the lemma, we first assume that $\sqrt{n}(\tilde{\theta} - \theta) = O_P(1)$. Then the sequence of processes $\sqrt{n}(\mathbb{H}_{0,\tilde{\theta}} - H_{0,\theta})$ converges in $l^\infty[1, \sigma]$, by the preceding lemma and the delta method applied to the differentiable map $\theta \mapsto H_{0,\theta}$. At least this is true along subsequences along which $\sqrt{n}(\mathbb{H}_{0,\tilde{\theta}}(t) - H_{0,\theta}(t^{\theta/\tilde{\theta}}))$ and $\sqrt{n}(\tilde{\theta} - \theta)$ converge jointly in distribution. We can now apply Gill's Theorem 2 directly to obtain the desired result.

Finally, we extend this to sequences $\tilde{\theta}$ that satisfy only $\sqrt{n}(\tilde{\theta} - \theta) \geqslant \tilde{h}$ for some $\tilde{h} = O_P(1)$. We have the inequality, for $\theta_1 \leqslant \theta_2$,

$$0 \leqslant \sup_{1 \leqslant t \leqslant \sigma^{\theta_2}} (\hat{H}_{0,\theta_2} - \mathbb{H}_{0,\theta_2})(t) \leqslant \sup_{1 \leqslant t \leqslant \sigma^{\theta_1}} (\hat{H}_{0,\theta_1} - \mathbb{H}_{0,\theta_1})(t). \tag{7.1}$$

To see that this is valid, note that $s \mapsto \hat{H}_{0,\theta_1}(s^{\theta_1/\theta_2})$ is a concave function (for $\theta_1 \leqslant \theta_2$) and a majorant of the function $s \mapsto \mathbb{H}_{0,\theta_1}(s^{\theta_1/\theta_2}) = \mathbb{H}_{0,\theta_2}(s)$. Thus it is bounded below by the least concave majorant $s \mapsto \hat{H}_{0,\theta_2}(s)$.

We apply this with $\theta_2 = \tilde{\theta}$ and $\theta_1 = \theta + \tilde{h}/\sqrt{n}$. Then the right-hand side converges to zero in probability by the preceding argument and the left-hand side is the variable of interest. □

We do not know if the second assertion of the preceding lemma is true for arbitrary estimators $\tilde{\theta}$. It appears to be not unlikely that slowly converging sequences $\tilde{\theta}$ may displace the 'observations' $e^{\tilde{\theta}T_i^*}$ too much to maintain good behaviour close to the point 1.

The preceding lemmas have been used in the proof of asymptotic normality of the maximum likelihood estimator $\hat{\theta}$. Once this is known, we can summarize the results concerning estimating $H_{0,\theta}$ and $H_{1,\theta}$ as follows.

Theorem 7.3. *Suppose that $H_{0,\theta}$ is continuous and strictly concave on its support $[1, \sigma_0]$, that $H_{0,\theta}$ and $H_{1,\theta}$ are continuously differentiable on the interval $[1, \sigma]$ and that the*

sequence $\sqrt{n}(\tilde{\theta} - \theta)$ *is asymptotically normal and asymptotically linear. Then the processes* $\sqrt{n}(\tilde{\theta} - \theta, \hat{H}_{0,\tilde{\theta}} - H_{0,\theta}, \mathbb{H}_{1,\tilde{\theta}} - H_{1,\theta})$ *converge in distribution to a tight Gaussian process in the space* $\mathbb{R} \times l^\infty[1, \sigma] \times l^\infty[1, \sigma]$.

Proof. By the preceding lemma the limit behaviour of $\hat{H}_{0,\tilde{\theta}}$ is the same as that of $\mathbb{H}_{0,\tilde{\theta}}$. We can decompose

$$\sqrt{n}(\mathbb{H}_{\delta,\tilde{\theta}} - H_{\delta,\theta})(t) = \sqrt{n}(\mathbb{F}^*_\delta - F^*_{\delta,\theta})\left(\frac{\log t}{\tilde{\theta}}\right) + \sqrt{n}\left(F^*_{\delta,\theta}\left(\frac{\log t}{\tilde{\theta}}\right) - F^*_{\delta,\theta}\left(\frac{\log t}{\theta}\right)\right).$$

The second term can be linearized in $\sqrt{n}(\tilde{\theta} - \theta)$ by the delta method. The first term can be handled by Lemma 9.1, after first noting that the empirical processes $\sqrt{n}(\mathbb{F}^*_\delta - F^*_{\delta,\theta})$ converge in the space $l^\infty[0, \infty]$ to tight Gaussian processes with uniformly continuous sample paths. These two sequences also converge jointly and jointly with the sequence $\sqrt{n}(\tilde{\theta} - \theta)$, as the marginals are asymptotically linear and satisfy the multivariate central limit theorem. □

8. Asymptotics for $\hat{h}_{0,\hat{\theta}}$

In this section we establish a uniform rate of convergence for the maximum likelihood estimator of the density $h_{0,\theta}$ and give its pointwise distribution. These results are needed in the proofs in the preceding sections, but are also of independent interest, because $\hat{h}_{0,\theta}$ is essentially the Grenander estimator.

In the following theorem we take σ to be equal to the end-point of the subdistribution $H_{0,\theta}$, i.e. the maximum of the support points of the conditional distribution of $e^{\theta T_i^*}$ given $\Delta_i = 0$ under the true parameter (θ, F, G).

Theorem 8.1. *Suppose that* $h_{0,\theta}$ *is continuously differentiable on the interval* $(1, \sigma)$ *with* $h'_{0,\theta}$ *bounded away from 0 (from above) and* $-\infty$ *and let* $\tilde{\theta}$ *be consistent for* θ. *Then for every* $x_n \to \infty$ *and* $\delta_n = n^{-1/3}(\log n)^{1/3}$,

$$\sup_{1 + x_n\delta_n \leqslant t < \sigma^{\tilde{\theta}/\theta}} |\hat{h}_{0,\tilde{\theta}}(t) - h_{0,\theta}(t^{\theta/\tilde{\theta}})t^{\theta/\tilde{\theta}-1}(\theta/\tilde{\theta})| = O_P(n^{-1/3}(\log n)^{1/3}).$$

If $\sqrt{n}(\tilde{\theta} - \theta) = O_P(1)$, *then this is also true for the supremum computed over* $1 \leqslant t < \sigma^{\tilde{\theta}/\theta}$.

Proof. To simplify notation, let $H_\eta(s) = H_{0,\theta}(s^{\theta/\eta})$ be the subdistribution function of the variables $e^{\eta T_i^*}$ if $\Delta_i = 0$ (under the true parameter (θ, F, G), whence $H_\eta \neq H_{0,\eta}$ except for $\eta = \theta$), let $h_\eta(s) = h_{0,\theta}(s^{\theta/\eta})s^{\theta/\eta-1}\theta/\eta$ be the corresponding density, and let $\mathbb{H}_\eta(s) = \mathbb{H}_{0,\eta}(s)$ and $\mathbb{G}_\eta = \sqrt{n}(\mathbb{H}_\eta - H_\eta)$ be the corresponding empirical subdistribution function and empirical process. Furthermore, let $\hat{h}_\eta$ be the left derivative of the least concave majorant $\hat{H}_\eta$ of $\mathbb{H}_\eta$, so that $\hat{h}_{\tilde{\theta}} = \hat{h}_{0,\tilde{\theta}}$.

As in Groeneboom (1985) or van der Vaart and Wellner (1996, Figure 3.1), for every $t \in (1, \sigma^{\tilde{\theta}/\theta})$, a, $\delta_n > 0$

$$\tilde{h}_{\tilde{\theta}}(t) > a \Leftrightarrow \operatorname*{argmax}_{s:s\geqslant 0}\{\mathbb{H}_{\tilde{\theta}}(s) - as\} > t$$

$$\Leftrightarrow \operatorname*{argmax}_{h:\,h\geqslant -\delta_n^{-1}t}\{\mathbb{H}_{\tilde{\theta}}(t+\delta_n h) - a(t+\delta_n h)\} > 0$$

$$\Leftrightarrow \operatorname*{argmax}_{h:\,h\geqslant -\delta_n^{-1}t}\{\mathbb{G}_{\tilde{\theta}}(t+\delta_n h) + \sqrt{n}(H_{\tilde{\theta}}(t+\delta_n h) - H_{\tilde{\theta}}(t) - a\delta_n h)\} > 0.$$

Because $H_{\tilde{\theta}}(s)$ is equal to 0 and $H_{\tilde{\theta}}(\infty)$, respectively, for $s<1$ or $s>\max_{i:\Delta_i=0}e^{\tilde{\theta}T_i}$, we may actually restrict the first argmax to the domain $s\geqslant 1$ and $s\leqslant \sigma^{\tilde{\theta}/\theta}$, i.e. $h\in\delta_n^{-1}(1-t,\ \sigma^{\tilde{\theta}/\theta}-t)$. Choose $a=h_{\tilde{\theta}}(t)+x\delta_n$ with, for the moment, $x>0$ fixed. By a Taylor expansion, for some $0\leqslant\xi\leqslant 1$, which may depend on $(\tilde{\theta},\ \delta_n,\ h,\ t)$, and every t and $t+\delta_n h$ in $(1,\ \sigma^{\tilde{\theta}/\theta})$,

$$H_{\tilde{\theta}}(t+\delta_n h) - H_{\tilde{\theta}}(t) - a\delta_n h = \tfrac{1}{2}h'_{\tilde{\theta}}(t+\xi\delta_n h)\delta_n^2 h^2 - x\delta_n^2 h \begin{cases} \leqslant -c\delta_n^2 h^2 - x\delta_n^2 h, \\ \geqslant -d\delta_n^2 h^2 - x\delta_n^2 h, \end{cases}$$

for certain $c,\ d>0$ independent of $\delta_n,\ t$ and h, and $\tilde{\theta}$ sufficiently close to θ, which will happen with probability tending to 1, since $\tilde{\theta}$ is consistent. Conclude that we can have $(\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}})(t) > x\delta_n$ only if, for any $h_0\in(\delta_n^{-1}(1-t),\ 0)$,

$$\sup_{h>0}(\mathbb{G}_{\tilde{\theta}}(t+\delta_n h) - \sqrt{n}\delta_n^2(ch^2+xh)) \geqslant \mathbb{G}_{\tilde{\theta}}(t+\delta_n h_0) - \sqrt{n}\delta_n^2(dh_0^2+xh_0).$$

Choose $h_0=-x/(2d)$ and note that $ch^2+xh\geqslant ch^2$ for $h\geqslant 0$. Then rearrange to find that

$$\mathrm{P}\left(\sup_{t\in(1+\delta_n x/(2d),\sigma^{\tilde{\theta}/\theta})}(\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}})(t) > x\delta_n\right)$$

$$\leqslant \mathrm{P}\left(\sup_{t\in(1,\sigma^{\tilde{\theta}/\theta}),\,h>0}(\mathbb{G}_{\tilde{\theta}}(t+\delta_n h) - \sqrt{n}\delta_n^2 ch^2 - \mathbb{G}_{\tilde{\theta}}(t+\delta_n h_0)) \geqslant \sqrt{n}\delta_n^2\frac{x^2}{4d}\right)$$

$$\leqslant \sum_{j=0}^{\infty}\mathrm{P}\left(\sup_{\substack{t\in(1,\sigma^{\tilde{\theta}/\theta})\\ j\leqslant h\leqslant j+1}}(\mathbb{G}_{\theta}((t+\delta_n h)^{\theta/\tilde{\theta}}) - \mathbb{G}_{\theta}((t+\delta_n h_0)^{\theta/\tilde{\theta}})) \geqslant \sqrt{n}\delta_n^2\left(cj^2+\frac{x^2}{4d}\right)\right).$$

Since $\tilde{\theta}$ is consistent, there is no loss of generality in assuming that $\theta/\tilde{\theta}$ is contained in an arbitrarily small neighbourhood $(1/r,\ r)$ of 1. Then defining the classes of functions

$$\mathscr{G}_{n,j} = \{1_{((t+\delta_n h_0)^\eta,(t+\delta_n h)^\eta]}:\ t\in(1,\ \sigma^r),\ j\leqslant h\leqslant j+1,\ \eta\in(1/r,\ r)\},$$

we can, using Markov's inequality, bound the preceding display further by a constant times

$$\sum_{j=0}^{\infty}\frac{\mathrm{E}\|\mathbb{G}_\theta\|_{\mathscr{G}_{n,j}}}{\sqrt{n}\delta_n^2(j^2+x^2)}. \tag{8.1}$$

Let $\lesssim$ denote inequality up to a constant. For a typical function $g\in\mathscr{G}_{n,j}$ we have

$$H_\theta g^2 = H_\theta((t+\delta_n h)^\eta) - H_\theta((t+\delta_n h_0)^\eta) \lesssim \delta_n(j+x)(t+\delta_n h)^{\eta-1} \lesssim \delta_n(j+x)^r.$$

The class $\mathscr{G}_{n,j}$ is uniformly bounded and contained in the class $\{1_{(a,b]}: a<b\}$, which is Vapnik–Chervonenkis of index 3. This implies first that $\mathrm{E}\|\mathbb{G}_\theta\|_{\mathscr{G}_{n,j}}$ is uniformly bounded, by, for example, Theorems 2.14.1 and 2.6.4 of van der Vaart and Wellner (1996). Second, for $0<\varepsilon<1$, by, for example, Example 2.5.4 in van der Vaart and Wellner (1996) and a simple argument,

$$N_{[\,]}(\varepsilon, \mathscr{G}_{n,j}, L_2(H_\theta)) \lesssim \left(\frac{1}{\varepsilon}\right)^4.$$

Consequently, by Lemma 3.4.2 of van der Vaart and Wellner (1996), there exists a constant C such that

$$\mathrm{E}\|\mathbb{G}_\theta\|_{\mathscr{G}_{n,j}} \lesssim J(C\sqrt{\delta_n(j+x)^r}) + \frac{J^2(C\sqrt{\delta_n(j+x)^r})}{\delta_n(j+x)^r\sqrt{n}},$$

where $J(\delta)$ is the entropy-with-bracketing integral, defined by

$$J(\delta) = \int_0^\delta \sqrt{1+\log N_{[\,]}(\varepsilon, \mathscr{G}_{n,j}, L_2(H_\theta))}\,\mathrm{d}\varepsilon.$$

For $\delta \lesssim \frac{1}{2}$, we have that $J(\delta) \lesssim \delta\sqrt{\log 1/\delta}$. Conclude that we can bound (8.1) up to a constant by

$$\sum_{j=0}^\infty \frac{1}{\sqrt{n}\delta_n^2(j^2+x^2)}\left(\left[J(\sqrt{\delta_n(j+x)^r}) + \frac{J^2(\sqrt{\delta_n(j+x)^r})}{\delta_n(j+x)^r\sqrt{n}}\right] \wedge 1\right)$$

$$\lesssim \sum_{\substack{j=0\\(j+x)^r\lesssim 1/\delta_n}}^\infty \left(\frac{\sqrt{|\log(\delta_n(j+x))|}}{\sqrt{n}\delta_n^{3/2}(j+x)^{2-r/2}} + \frac{|\log(\delta_n(j+x))|}{n\delta_n^2(j+x)^2}\right) + \sum_{\substack{j=0\\(j+x)^r\gtrsim 1/\delta_n}}^\infty \frac{1}{\sqrt{n}\delta_n^2(j+x)^2}.$$

The last term on the right-hand side can be bounded by

$$\sum_{j=0}^\infty \frac{1}{\sqrt{n}\delta_n^{3/2}(j+x)^{2-r/2}}.$$

For $\delta_n = (n/\log n)^{-1/3}$ all terms converge to 0 as $x = x_n \to \infty$.

This, combined with a similar argument for the lower tail of the distribution of $\sup_t(\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}})(t)$, shows that for every $\varepsilon>0$ there exists x such that, for all sufficiently large n,

$$\mathrm{P}\left(\sup_{1+x\delta_n\le t<\sigma^{\tilde{\theta}/\theta}} |\hat{h}_{\tilde{\theta}}(t) - h_{\tilde{\theta}}(t)| > x\delta_n\right) < \varepsilon.$$

This implies the first assertion of the theorem.

Because $F_0\{0\} = 0$ by assumption, we have $h_\theta(1) = H_\theta\{1\} = h_\theta(1+)$ by the last assertion of Lemma 1.1. Therefore the function $h_{\tilde{\theta}}$ is Lipschitz on the interval $[1, 1+x\delta_n]$.

Combining this with the monotonicity of $\hat{h}_{\tilde{\theta}}$, we see that there exists a constant C such that, for every $t \in [1, 1 + x\delta_n]$,

$$(\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}})(1 + \delta_n x) - C\delta_n x \leqslant (\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}})(t) \leqslant (\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}})(1) + C\delta_n x.$$

Here, by Lemma 7.2, up to $O_P(n^{-1/2})$, the variable $\hat{h}_{\tilde{\theta}}(1) = \hat{H}_{\tilde{\theta}}(1)$ is equal to $H_{\tilde{\theta}}(1) = H_\theta\{1\} = h_\theta(1) = h_{\tilde{\theta}}(1)(\tilde{\theta}/\theta)$. It follows that

$$\sup_{1 \leqslant t < \sigma^{\tilde{\theta}/\theta}} |\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}}|(t) \leqslant \sup_{1 + \delta_n x \leqslant t < \sigma^{\tilde{\theta}/\theta}} |\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}}|(t) + C\delta_n x + O_P(n^{-1/2} + |\tilde{\theta} - \theta|).$$

We conclude that for every $\varepsilon > 0$ there exists x such that the left-hand side is bounded by a fixed multiple of $x\delta_n$ with probability at least $1 - \varepsilon$. This proves the theorem. □

The properties of our estimator $\hat{h}_{0,\theta}$ are closely related to the properties of the Grenander estimator of a monotone density. By simplifying the preceding proof it can be shown that the Grenander estimator $\hat{h}$ of a monotone density h on $(0, \infty)$ satisfies, for every $x_n \to \infty$,

$$\sup_{x_n\delta_n < t < \sigma} |\hat{h} - h|(t) = O_P(\delta_n),$$

for $\delta_n = n^{-1/3}(\log n)^{1/3}$, under the condition that h possesses a derivative that is bounded, strictly negative and bounded away from zero. The fact that we need to restrict the range of the uniform norm to $t > x_n\delta_n$ is consistent with the known fact that the Grenander estimator is not consistent at 0. In the situation of the present paper, the distribution starts with a point mass at 1, and a similar problem at the left boundary of the support of h does not occur, as is argued explicitly at the end of the proof of the last theorem.

In the following corollaries we implicitly assume the same conditions as in the preceding theorem.

Corollary 8.2. *If* $\sqrt{n}(\tilde{\theta} - \theta) = O_P(1)$, *then*

$$\sup_{1 < t < \sigma \wedge \sigma^{\tilde{\theta}/\theta}} |\hat{h}_{0,\tilde{\theta}} - h_{0,\theta}|(t) = O_P(n^{-1/3}(\log n)^{1/3}).$$

Proof. This is a consequence of the preceding theorem and the differentiability of $t \mapsto h_{0,\theta}(t)$ on the interval $(1, \sigma)$. □

Corollary 8.3. *Let* $\hat{\sigma}_0 = 1$ *and let* $\hat{\sigma}_1 < \ldots < \hat{\sigma}_{K_n}$ *be the points in* $[1, \sigma^{\tilde{\theta}/\theta}]$ *where the least concave majorant* $\hat{H}_{0,\tilde{\theta}}$ *changes direction. Then*

$$\max_{1 \leqslant i \leqslant K_n} \hat{\sigma}_i - \hat{\sigma}_{i-1} = O_P(n^{-1/3}(\log n)^{1/3}).$$

Proof. We adopt the notation of the preceding proofs. Let $x_n \to \infty$ be arbitrary, let $\overline{\sigma}_0 = 1 + x_n\delta_n$, and let $\overline{\sigma}_1 < \ldots < \overline{\sigma}_{L_n}$ be the points $\hat{\sigma}_i$ that are contained in $(1 + x_n\delta_n, e^{\tilde{\theta}\tau}]$. Then

$$\max_i(\hat{\sigma}_i - \hat{\sigma}_{i-1}) \leqslant x_n\delta_n + \max_i(\overline{\sigma}_i - \overline{\sigma}_{i-1}).$$

Hence, because $x_n \to \infty$ is arbitrary, it suffices to prove the corresponding property of the $\overline{\sigma}_i$ instead of the $\hat{\sigma}_i$.

If the maximum between the points $\overline{\sigma}_i$ is larger than δ_n, then there is a subinterval of $(\overline{\sigma}_0, \sigma^{\tilde{\theta}/\theta})$ of length at least δ_n on which $\hat{h}_{\tilde{\theta}}$ is constant. Hence

$$\mathrm{P}\left(\max_{1 \leqslant i \leqslant K_n} \overline{\sigma}_i - \overline{\sigma}_{i-1} \geqslant \delta_n\right) \leqslant \mathrm{P}\left(\exists s,\ t \in (\overline{\sigma}_0,\ \sigma^{\tilde{\theta}/\theta})\colon |s-t| \geqslant \delta_n,\ \hat{h}_{\tilde{\theta}}(s) = \hat{h}_{\tilde{\theta}}(t)\right).$$

Since $h'_\theta(s)$ is bounded away from zero on $(1, \sigma)$, h_θ is bounded above, and $\tilde{\theta}$ is consistent, $h'_{\tilde{\theta}}$ is bounded away from zero on $(1, \sigma^{\tilde{\theta}/\theta})$ with probability tending to 1. Therefore, there exists a constant $c > 0$ such that, with probability tending to 1, $|h_{\tilde{\theta}}(s) - h_{\tilde{\theta}}(t)| \geqslant c|s-t|$ for every $s,\ t \in (1, \sigma^{\tilde{\theta}/\theta})$. Hence on the event in the right-hand side of the preceding display we have, with probability tending to 1, that there exist $s,\ t \in (\overline{\sigma}_0, \sigma^{\tilde{\theta}/\theta})$ such that

$$c\delta_n \leqslant |h_{\tilde{\theta}}(s) - h_{\tilde{\theta}}(t)| = |\hat{h}_{\tilde{\theta}}(s) - \hat{h}_{\tilde{\theta}}(t) - (h_{\tilde{\theta}}(s) - h_{\tilde{\theta}}(t))|.$$

Thus the probability of this event is bounded above by

$$\mathrm{P}\left(2 \sup_{\overline{\sigma}_0 < t < \sigma^{\tilde{\theta}/\theta}} |\hat{h}_{\tilde{\theta}} - h_{\tilde{\theta}}|(t) \geqslant c\delta_n\right).$$

An application of Theorem 8.1 concludes the proof. □

Corollary 8.4. *Let* $\hat{\sigma}_0 = 1$ *and let* $\hat{\sigma}_1 < \ldots < \hat{\sigma}_{K_n}$ be the points in $[1, \sigma^{\tilde{q}/\theta}]$ *where the least concave majorant* $\hat{H}_{0,\tilde{\theta}}$ *changes direction, and let* $m_{n,i}$ *be the number of variables* $\mathrm{e}^{\tilde{\theta}T_i^*}$ *with* $\Delta_i = 0$ *that fall in* $(\hat{\sigma}_{i-1}, \hat{\sigma}_i]$. *Then* $\max_{1 \leqslant i \leqslant K_n} m_{n,i} = O_P(n^{2/3}(\log n)^{1/3})$.

Proof. We adopt the notation used in the proof of the preceding corollary. Let $\varepsilon > 0$ be small and fix $c > \sup\{h_\eta(t)\colon t \in (1, \sigma^{\eta/\theta}],\ |\eta - \theta| < \varepsilon\}$. By Corollary 8.3, $\max_i(\overline{\sigma}_i - \overline{\sigma}_{i-1})$ is bounded by $x_n\delta_n/c$ with probability tending to 1, if $x_n \to \infty$. If $\max_{1 \leqslant i \leqslant K_n} m_{n,i} \geqslant nx_n\delta_n$, then there exists an interval $(\overline{\sigma}_{i-1}, \overline{\sigma}_i]$ on which $\mathbb{H}_{0,\tilde{\theta}}$ increases by at least $x_n\delta_n$. On this interval $\hat{h}_{0,\tilde{\theta}}$ is at least $x_n\delta_n/(\overline{\sigma}_i - \overline{\sigma}_{i-1})$. With probability tending to 1 this is at least c. Since $\hat{h}_{0,\tilde{\theta}}$ is uniformly consistent for $h_{0,\theta}$ on $[1 + x_n\delta_n, \sigma^{\tilde{\theta}/\theta}]$; this can happen only with probability tending to 0 by the definition of c.

This leaves out the intervals $(\hat{\sigma}_{i-1}, \hat{\sigma}_i]$ contained in $[1, 1 + x_n\delta_n]$. However, the number of $\mathrm{e}^{\tilde{\theta}T_i^*}$ falling in the interval $(1, 1 + x_n\delta_n]$ is of order $nH_{0,\theta}(1, (1 + x_n\delta_n)^{\theta/\tilde{\theta}}] = O_P(nx_n\delta_n)$, which is of the same order as the upper bound over the other intervals if x_n is large but fixed. □

Theorem 8.5. *Suppose that* $h_{0,\theta}$ *is continuously differentiable in a neighbourhood of* t *with* $h'_{0,\theta}(t) < 0$. *Then for any random sequences* $\tilde{t} = t + o_P(n^{-1/3})$ *and* $\tilde{\theta} = \theta + o_P(n^{-1/3})$,

$$n^{1/3}(\hat{h}_{0,\tilde{\theta}}(\tilde{t}) - h_{0,\tilde{\theta}}(t)) \rightsquigarrow |4h'_{0,\theta}(t)h_{0,\theta}(t)|^{1/3} \operatorname*{argmax}_{h \in \mathbb{R}} \{Z(h) - h^2\},$$

where Z *is a standard Brownian motion. Moreover,* $n^{1/3}(\hat{h}_{0,\tilde{\theta}}(\tilde{t}_n) - \hat{h}_{0,\tilde{\theta}}(t)) \xrightarrow{\mathrm{P}} 0$.

Proof. We use the notation of the proof of Theorem 8.1. By the first part of this proof

$$n^{1/3}(\hat{h}_{\tilde{\theta}}(\tilde{t}) - h_{\tilde{\theta}}(\tilde{t})) \leqslant x \Leftrightarrow \operatorname*{argmax}_{h: h \geqslant -\delta_n^{-1}\tilde{t}} Z_n(h) \leqslant 0,$$

where $h \mapsto Z_n(h)$ is the stochastic process

$$Z_n(h) = n^{1/6}(\mathbb{G}_\theta((\tilde{t} + \delta_n h)^{\theta/\tilde{\theta}}) - \mathbb{G}_\theta(\tilde{t}^{\theta/\tilde{\theta}})) + n^{2/3}(H_{\tilde{\theta}}(\tilde{t} + \delta_n h) - H_{\tilde{\theta}}(\tilde{t}) - h_{\tilde{\theta}}(\tilde{t})\delta_n h) - xh.$$

By the empirical central limit theorem (see Theorem 2.11.22 in van der Vaart and Wellner 1996), Lemma 9.1, and the twice continuous differentiability of $t \mapsto H_\theta(t)$, the sequence Z_n converges for every fixed M in $l^\infty[-M, M]$ to the process

$$\sqrt{h_\theta(t)}Z(h) + \tfrac{1}{2}h'_\theta(t)h^2 - xh.$$

We show below that the argmax $\hat{h}_n$ of the processes Z_n are bounded in probability whenever the diameters of the ranges of $\tilde{t}_n$ are of order $O_P(n^{-1/3})$. It then follows from the continuous mapping theorem for the argmax functional (van der Vaart and Wellner 1996, Theorem 3.2.2) that $\hat{h}_n \rightsquigarrow \hat{h}$, for $\hat{h}$ the argmax of the process $h \mapsto \sqrt{h_\theta(t)}Z(h) + \frac{1}{2}h'_\theta(t)h^2 - xh$. Using rescaling properties of Brownian motion, the probability $\mathrm{P}(\hat{h} \leqslant 0)$ can be rewritten as (cf. van der Vaart and Wellner 1996, Problem 3.2.5)

$$\mathrm{P}(\hat{h} \leqslant 0) = \mathrm{P}(|4h'_{0,\theta}(t)h_{0,\theta}(t)|^{1/3} \operatorname*{argmax}_{h \in \mathbb{R}}\{Z(h) - h^2\} \leqslant x).$$

Since $h_{\tilde{\theta}}(t) = h_{0,\theta}(t^{\theta/\tilde{\theta}})t^{\theta/\tilde{\theta}-1}(\theta/\tilde{\theta})$, $\tilde{\theta} = \theta + o_P(\delta_n)$ and $h_{0,\theta}$ is differentiable, this yields the first statement of the theorem.

Actually, our proof shows that the same limit law is obtained for $n^{1/3}(\hat{h}_{\tilde{\theta}}(\tilde{t}) - h_{\tilde{\theta}}(\tilde{t}))$ for any sequence $\tilde{t}_n \xrightarrow{\mathrm{P}} t$ with the special property described above. For sequences $\tilde{t}_n = t + o_P(\delta_n)$ that converge to t fast, we also have that the limit processes Z constructed in the preceding argument can be coupled and be taken to be equal to the process $Z(h)$ obtained for $\tilde{t}_n \equiv t$. This follows from the fact that in this case, by Lemma 9.1,

$$\sup_{|h| \leqslant M} n^{1/6}|\mathbb{G}_\theta((\tilde{t} + \delta_n h)^{\theta/\tilde{\theta}}) - \mathbb{G}_\theta((t + \delta_n h)^{\theta/\tilde{\theta}})| \xrightarrow{\mathrm{P}} 0.$$

Thus, for $\tilde{t}_n = t + o_P(\delta_n)$, we find

$$\mathrm{P}(n^{1/3}(\hat{h}_{\tilde{\theta}}(\tilde{t}) - h_{\tilde{\theta}}(\tilde{t})) \leqslant x,\ n^{1/3}(\hat{h}_{\tilde{\theta}}(t) - h_{\tilde{\theta}}(t)) > x) \to \mathrm{P}(\hat{h} \leqslant 0,\ \hat{h} > 0),$$

and similarly for the inequalities $\leqslant$ and $>$ interchanged. This implies the second statement of the theorem. (If $\mathrm{P}(X \leqslant x,\ Y > x) = 0$ for every x, then $X \geqslant Y$ almost surely.)

Finally, we show that $\hat{h}_n = O_P(1)$. For this purpose we apply a general theorem on rates of convergence of M-estimators. Specifically, we apply Theorem 5.55 of van der Vaart (1998) (cf. Theorem 3.2 of Murphy and Van der Vaart 1999), which allows for nuisance parameters, with the choices $\eta = (t', \delta, \theta')$ and $\hat{\eta} = (\tilde{t}, \delta_n, \tilde{\theta})$ and

$$m_{h,\eta} = 1_{[0,(t'+h)^{\theta/\theta'}]} - 1_{[0,(t')^{\theta/\theta'}]} - h_{\theta'}(t')h - xh\delta.$$

Then $\delta_n\hat{h}_n$ maximizes $\mathbb{P}_n m_{h,\hat{\eta}}$ for $\mathbb{P}_n$ the empirical subdistribution of the points $\mathrm{e}^{\theta T_i^*}$ with

$\Delta_i = 0$. For h sufficiently close to 0 and $\eta = (t', \delta, \theta')$ sufficiently close to $\eta_0 := (t, 0, \theta)$ we have, by the concavity of $t \mapsto H_\theta(t)$, for some constants C, C_1, $C_2 > 0$,

$$H_\theta(m_{h,\eta} - m_{0,\eta}) \leqslant -Ch^2 + |xh\delta| \leqslant -C_1 h^2 + C_2\delta^2.$$

Furthermore, we have $\mathbb{G}_\theta(m_{h,\eta} - m_{0,\eta}) = \mathbb{G}_\theta(1_{[0,(t'+h)^{\theta/\theta'}]} - 1_{[0,(t')^{\theta/\theta'}]})$ and the functions $1_{[0,(t'+h)^{\theta/\theta'}]} - 1_{[0,(t')^{\theta/\theta'}]}$ satisfy

$$H_\theta(1_{[0,(t'+h)^{\theta/\theta'}]} - 1_{[0,(t')^{\theta/\theta'}]})^2 \leqslant C_4|h|.$$

For $|h| < \delta$, $|t' - t| \leqslant M\delta$ and $|\theta' - \theta| \leqslant M\delta$, these functions are indicators of cells with end-points contained in an interval of length proportional to δ. Therefore,

$$\mathrm{N}_{[\,]}(\varepsilon, \{1_{[0,(t'+h)^{\theta'/\theta}]} - 1_{[0,(t')^{\theta'/\theta}]} : |h| < \delta, |t' - t| \leqslant M\delta, |\theta' - \theta| \leqslant M\delta\}, L_2(H_\theta)) \lesssim \frac{\delta^2}{\varepsilon^4}.$$

It follows by the maximal inequality given by Lemma 3.4.2 of van der Vaart and Wellner (1996) that

$$\mathrm{E} \sup_{|h|<\delta, |t'-t|\leqslant M\delta, |\theta'-\theta|<M\delta} |\mathbb{G}_\theta(m_{h,\eta} - m_{0,\eta})| \lesssim J(\delta)\left(1 + \frac{J(\delta)}{\delta^2\sqrt{n}}\right),$$

for

$$J(\delta) = \int_0^{\sqrt{\delta}} \sqrt{\log\left(1 + \left(\frac{\delta^2}{\varepsilon^4}\right)\right)}\,\mathrm{d}\varepsilon \lesssim \sqrt{\delta}.$$

Therefore, by Theorem 5.55 of van der Vaart (1998), the rate of convergence of $\delta_n\hat{h}_n$ to 0 is $O_P(\delta_n)$, provided that $\delta_n\hat{h}_n$ is, with probability tending to 1, in the neighbourhood of 0 used in the preceding estimates. The latter, and even that $\delta_n\hat{h}_n \xrightarrow{\mathrm{P}} 0$, can be proved by a direct argument using the Glivenko–Cantelli theorem. □

9. Proof of Theorem 5.1

Let σ be strictly bigger than $\mathrm{e}^{\theta\tau}$ (for θ the true value) and such that $((1 - F)(1 - G))(\log\sigma/\theta) > 0$. Since $\hat{\theta}$ is consistent, we can assume without loss of generality that $\mathrm{e}^{\hat{\theta}t} \leqslant \sigma$ with probability 1 for every $t \leqslant \tau$. Then $\hat{\Lambda}_F$ and Λ_F on the interval $[0, \tau]$ depend on the values of $(H_{0,\theta}, H_{1,\theta}, h_{0,\theta})$ and their estimators on the interval $[1, \sigma]$ only.

It follows from Theorems 4.1 and 7.3 that $\sqrt{n}(\hat{\theta} - \theta, \hat{H}_{0,\hat{\theta}} - H_{0,\theta}, \hat{H}_{1,\hat{\theta}} - H_{1,\theta})$ converges in distribution to a tight Gaussian variable in the space $\mathbb{R} \times l^\infty[1, \sigma] \times l^\infty[1, \sigma]$, and it follows from Corollary 8.2 that $\hat{h}_{0,\hat{\theta}} - h_{0,\theta}$ converges in probability to 0 uniformly on $[1, \sigma]$ (at the rate of almost $n^{-1/3}$).

The maximum likelihood estimator can be written as

$$\hat{\Lambda}_F(t) = \phi(\hat{H}_{0,\hat{\theta}} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, \hat{H}_{1,\hat{\theta}})(\mathrm{e}^{\hat{\theta}t}),$$

for

$$\phi(H_0, H_1)(t) = \int_{[1,t]} \frac{\mathrm{d}H_1}{1 - H_{1-} - H_{0-}}.$$

Thus $\hat{\Lambda}_F(t)$ is formed in two steps: first, the composition of the stochastic process $t \mapsto \phi(\hat{H}_{0,\hat{\theta}} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, \hat{H}_{1,\hat{\theta}})(t)$ indexed by $t \in [1, \sigma]$; and second, the change of scale $t \mapsto \mathrm{e}^{\hat{\theta}t}$. We analyse these two steps separately by decomposing

$$\begin{aligned}(\hat{\Lambda}_F - \Lambda_F)(t) = {} & (\phi(\hat{H}_{0,\hat{\theta}} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, \hat{H}_{1,\hat{\theta}}) - \phi(H_{0,\theta} - \mathrm{id}h_{0,\theta}, H_{1,\theta}))(\mathrm{e}^{\hat{\theta}t}) \\ & + \phi(H_{0,\theta} - \mathrm{id}h_{0,\theta}, H_{1,\theta})(\mathrm{e}^{\hat{\theta}t}) - \phi(H_{0,\theta} - \mathrm{id}h_{0,\theta}, H_{1,\theta})(\mathrm{e}^{\theta t}).\end{aligned} \tag{9.1}$$

The second term can be linearized in $\hat{\theta} - \theta$ by an application of the delta method for Euclidean variables. The first term concerns the processes $t \mapsto \phi(\hat{H}_{0,\hat{\theta}} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, \hat{H}_{1,\hat{\theta}}) - \phi(H_{0,\theta} - \mathrm{id}h_{0,\theta}, H_{1,\theta})$ evaluated on a random time-scale. By the following lemma, the limit distribution of this term remains the same if the random time is replaced by the fixed time $\mathrm{e}^{\theta t}$.

Let S and T be arbitrary sets, and define $l^\infty(S)$ as the Banach space of all bounded functions $z: S \mapsto \mathbb{R}$ equipped with the uniform norm.

Lemma 9.1. *Suppose that Z_n are random elements in the space $l^\infty(T)$ such that $Z_n \rightsquigarrow Z$ for a tight, Borel measurable Gaussian process Z. If $\hat{g}_n: S \mapsto T$ are random maps such that $\sup_{s\in S} d_Z(\hat{g}_n(s), g(s)) \xrightarrow{\mathrm{P}} 0$ for a fixed map $g: S \mapsto T$ and $d_Z^2(t_1, t_2)$ the second moment of $Z(t_1) - Z(t_2)$, then $Z_n \circ \hat{g}_n - Z_n \circ g \rightsquigarrow 0$ in $l^\infty(S)$.*

This lemma is a consequence of the fact that the sample paths of a tight Gaussian process Z in $l^\infty(T)$ are automatically uniformly continuous relative to the second-moment semi-metric d_Z. The lemma is a more abstract version of Lemma 3.3.5 in van der Vaart and Wellner (1996) and can be formally proved along the same lines. Instead of the semi-metric d_Z, we may use any semi-metric d for which T is totally bounded and such that the sample paths of Z are uniformly continuous relative to d.

We have that $\sup_{0\leqslant t\leqslant\tau]}|\mathrm{e}^{\hat{\theta}t} - \mathrm{e}^{\theta t}| \xrightarrow{\mathrm{P}} 0$. Thus we handle the first term of (9.1) by proving the weak convergence of $\sqrt{n}$ times the processes $\phi(\hat{H}_{0,\hat{\theta}} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, \hat{H}_{1,\hat{\theta}}) - \phi(H_{0,\theta} - \mathrm{id}h_{0,\theta}, H_{1,\theta})$ in $l^\infty[1, \sigma]$, and showing that the second-moment metric of the limit process is continuous relative to the Euclidean distance. To take the special properties of $\hat{h}_{0,\hat{\theta}}$ into account, we decompose these processes as

$$\begin{aligned}& \phi(\hat{H}_{0,\hat{\theta}} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, \hat{H}_{1,\hat{\theta}}) - \phi(H_{0,\theta} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, H_{1,\theta}) \\ & \quad + \phi(H_{0,\theta} - \mathrm{id}\hat{h}_{0,\hat{\theta}}, H_{1,\theta}) - \phi(H_{0,\theta} - \mathrm{id}h_{0,\theta}, H_{1,\theta})\end{aligned} \tag{9.2}$$

We linearize the two terms separately by an extension of the functional delta method. Let BV and $\mathbb{D}$ be the set of all functions $z: [1, \sigma] \mapsto \mathbb{R}$ of bounded variation and the set of functions that are left- or right-continuous with limits from the left and right everywhere, respectively, and let BV_1 be the unit ball in the first space. We equip $\mathbb{D}$ with the uniform norm. The map $\phi: \mathbb{D} \times \mathrm{BV}_1 \subset \mathbb{D} \times \mathbb{D} \mapsto \mathbb{D}$ is Hadamard differentiable at every point $(H_0, H_1) \in \mathbb{D}_\phi$ such that $H_0 \in \mathrm{BV}$ if restricted to the domain $\mathbb{D}_\phi$ of points (H_0, H_1) such that $\int \mathrm{d}|H_1| \leqslant 1$ and $1 - H_0(\sigma) - H_1(\sigma) > \varepsilon$ for some $\varepsilon > 0$. This follows with the help of the chain rule and

standard results (see Gill 1989; or van der Vaart and Wellner 1996) from Hadamard calculus, since the map can be decomposed as

$$(H_0,\, H_1) \mapsto \left(\frac{1}{1 - H_{0-} - H_{1-}},\, H_1\right) \mapsto \int_{[0,\cdot]} \frac{\mathrm{d}H_1}{1 - H_{0-} - H_{1-}}.$$

The function $1 - H_{1,\theta}(y) - H_{0,\theta}(y) + yh_{0,\theta}(y+) = 1 - F_0(\log y/\theta) - F_1(\log y/\theta)$ is bounded away from zero on $[1, \sigma]$ by assumption. Since $\hat{H}_{0,\hat{\theta}}$, $\hat{H}_{1,\hat{\theta}}$ and $\hat{h}_{0,\hat{\theta}}$ are consistent, the same is true with probability tending to 1 for the functions obtained by substituting these estimators. Therefore, the Hadamard differentiability of ϕ is sufficient to infer that, for $h = h_{0,\theta}$,

$$\begin{aligned}\phi(\hat{H}_{0,\hat{\theta}} - \mathrm{id}h,\, \hat{H}_{1,\hat{\theta}}) - \phi(H_{0,\theta} - \mathrm{id}h,\, H_{1,\theta}) \\ = \phi'_{H_{0,\theta} - \mathrm{id}h, H_{1,\theta}}(\hat{H}_{0,\hat{\theta}} - H_{0,\theta},\, \hat{H}_{1,\hat{\theta}} - H_{1,\theta}) + o_P\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

where ϕ'_{H_0, H_1} is the derivative of ϕ at (H_0, H_1). This is not enough to handle the first term on the right-hand side of (9.2), because there h is taken equal to the random variables $\hat{h}_{0,\hat{\theta}}$. However, the preceding display remains valid if h is replaced by $\hat{h}_{0,\hat{\theta}}$, where we may evaluate the derivative on the right-hand side at the uniform limit in probability $h_{0,\theta}$ of $\hat{h}_{0,\hat{\theta}}$. To see this, we must make the delta method 'locally uniform'.

Suppose that $\phi : \mathbb{D}_\phi \subset \mathbb{D}_1 \times \mathbb{D}_2 \mapsto \mathbb{D}_3$ is a map defined on a subset $\mathbb{D}_\phi$ of a product of two normed spaces with values in a third normed space. Call ϕ *Hadamard differentiable at* (A, B) *locally uniformly in* A if for all converging sequences $A_t \to A$, $a_t \to a$ and $b_t \to b$ as $t \downarrow 0$ such that $(A_t + ta_t,\, B + tb_t) \in \mathbb{D}_\phi$ and $(A_t, B) \in \mathbb{D}_\phi$ for every t,

$$\frac{\phi(A_t + ta_t,\, B + tb_t) - \phi(A_t, B)}{t} \to \phi'_{A,B}(a, b),$$

for a continuous, linear map $\phi'_{A,B} : \mathbb{D}_1 \times \mathbb{D}_2 \mapsto \mathbb{D}_3$. Then we have the following extension of the delta method theorem (cf. van der Vaart and Wellner 1996, Section 3.9.1).

Lemma 9.2. *Suppose* $\phi : \mathbb{D}_\phi \subset \mathbb{D}_1 \times \mathbb{D}_2 \mapsto \mathbb{D}_3$ *is Hadamard differentiable at* (A, B) *locally uniformly in* A. *If* $(X_n, Y_n, \hat{A}_n)$ *are random elements such that* $(X_n, Y_n) \in \mathbb{D}_\phi$ *and* $(\hat{A}_n, B) \in \mathbb{D}_\phi$ *and such that* $\sqrt{n}(X_n - \hat{A}_n,\, Y_n - B)$ *converges in distribution in* $\mathbb{D}_1 \times \mathbb{D}_2$ *to a tight limit* (X, Y) *and* $\hat{A}_n \xrightarrow{\mathrm{P}} A$, *then*

$$\phi(X_n, Y_n) - \phi(\hat{A}_n, B) = \phi'_{A,B}(X_n - \hat{A}_n,\, Y_n - B) + o_P(n^{-1/2}).$$

Proof. Define maps g_n by $g_n(x, y, a) = \sqrt{n}(\phi(a + n^{-1/2}x,\, B + n^{-1/2}y) - \phi(a, B)) - \phi'_{A,B}(x, y)$. Then $g_n(x_n, y_n, a_n) \to 0$ for all converging sequences $x_n \to x$, $y_n \to y$ and $a_n \to A$. Consequently, by the extended continuous mapping theorem (van der Vaart and Wellner 1996, Theorem 1.11.1) $g_n(\sqrt{n}(X_n - \hat{A}_n), \sqrt{n}(Y_n - B), \hat{A}_n) \rightsquigarrow 0$. Since convergence in probability and convergence in distribution to a degenerate limit are the same, this is the assertion of the lemma. □

If we establish the Hadamard differentiability, locally uniform in the argument H_0, of the map $(H_0, H_1) \mapsto \phi(H_0, H_1)$ as defined previously, then the desired linearization result follows. We can achieve this by the same method as for the proof of ordinary Hadamard differentiability of ϕ (cf. Gill 1989; or Lemmas 3.9.25 and 3.9.17 in van der Vaart and Wellner 1996). Here we may use an appropriate version of the chain rule, which remains valid under the extension to locally uniform differentiability. The following lemma states the Hadamard differentiability of the hardest constituent of ϕ, the Wilcoxon map $(A, B) \mapsto \int A \, \mathrm{d}B$.

Lemma 9.3. *The map* $\phi : \mathrm{BV}_1 \times \mathbb{D} \subset \mathbb{D} \times \mathbb{D} \mapsto \mathbb{D}$ *given by* $\phi(A, B) = \int_{[1,\cdot]} A \, \mathrm{d}B$ (*defined by partial integration if necessary*) *is Hadamard differentiable at every* $(A, B) \in \mathrm{BV}_1 \times \mathrm{BV}$ *locally uniformly in* A.

The lemma can be proved by a minor adaptation of Lemma 3.9.17 in van der Vaart and Wellner (1996).

Along the same lines, we can also introduce the concept of Hadamard differentiability locally uniform in both A and B. This concept would be close to continuous differentiability, which is slightly stronger (cf. van der Vaart and Wellner 1996, Lemma 3.9.7). This concept is useful for the analysis of the map $(H_0, H_1) \mapsto 1/(1 - H_0 - H_1)$, but too restrictive in the case of the Wilcoxon map, which is only partially locally uniformly differentiable.

Thus, the first term on the right-hand side of (9.2) can be linearized in $(\hat{H}_{0,\hat{\theta}} - H_{0,\theta}, \hat{H}_{1,\hat{\theta}} - H_{1,\theta})$, and to find its limit law we may set $\hat{h}_{0,\hat{\theta}}$ equal to its limit $h_{0,\theta}$.

The second term on the right of (9.2) can be written $\psi(\hat{h}_{0,\theta}, \hat{H}_{0,\hat{\theta}}) - \psi(\hat{h}_{0,\theta}, H_{0,\theta})$ for the map ψ defined by

$$\psi(h, H) = \int_{[1,\cdot]} \frac{-\mathrm{id}h_{1,\theta} \, \mathrm{d}H}{(1 - H_{0,\theta-} - H_{1,\theta-} + \mathrm{id}h_-)(1 - H_{0,\theta-} - H_{1,\theta-} + \mathrm{id}h_{0,\theta-})}.$$

(We use here the fact that $(\hat{h}_{0,\hat{\theta}} - h_{0,\theta})\mathrm{d}H_{1,\theta} = h_{1,\theta}\mathrm{d}(\hat{H}_{0,\hat{\theta}} - H_{0,\theta})$.) The map ψ is Hadamard differentiable at $(h_{0,\theta}, H_{0,\theta})$ locally uniformly in its first argument on the appropriate domain. (Actually, partial differentiability in its second argument, locally uniformly in its first argument would suffice.) Therefore, we can approximate this term by $\psi'_{h_{0,\theta}, H_{0,\theta}}(0, \hat{H}_{0,\theta} - H_{0,\theta})$.

The linear approximation to $\sqrt{n}(\hat{\Lambda}_F - \Lambda_F)$ obtained in this way is asymptotically Gaussian distributed. By the Hadamard differentiability of the product integral (Gill 1994; or van der Vaart and Wellner 1996, Lemma 3.9.30), this carries over into convergence in distribution of $\sqrt{n}(\hat{F} - F)$. Thus we have proved the theorem.

10. Proofs of Theorems 6.1 and 6.2

Because $\int |\mathrm{d}\hat{h}_{0,\hat{\theta}}| = O_P(1)$ and the differences between $\hat{H}_{0,\hat{\theta}}$ and $\hat{H}_{1,\hat{\theta}}$ and their limits are $O_P(n^{-1/2})$, we have, uniformly in $t \in [0, \tau]$,

$$\hat{\Lambda}_G(t) = -\int_{[1,e^{\hat{\theta}t}]} \frac{\mathrm{id}\,\mathrm{d}\hat{h}_{0,\hat{\theta}}}{1 - H_{0,\theta-} - H_{1,\theta} + \mathrm{id}\hat{h}_{0,\hat{\theta}-}} + O_P(n^{-1/2}).$$

Changing the integration range in the expression for $\Lambda_G(t)$ into the stochastic interval $[1, e^{\hat{\theta}t}]$ makes a difference of $O_P(n^{-1/2})$, by the delta method (since we assume that $h'_{0,\theta}$ exists and is bounded). Therefore, the difference between $\hat{\Lambda}_G(t)$ and $\Lambda_G(t)$ is up to $O_P(n^{-1/2})$ equal to $\int_{[1,e^{\hat{\theta}t}]}\phi(\mathrm{id}, \hat{h}_-)\mathrm{d}\hat{h} - \int_{[1,e^{\hat{\theta}t}]}\phi(\mathrm{id}, h)\mathrm{d}h$, for

$$\phi(u, v) = \frac{-u}{1 - H_{0,\theta}(u-) - H_{1,\theta}(u) + uv},$$

$h = h_{0,\theta}$, and $\hat{h} = \hat{h}_{0,\hat{\theta}}$. It can be decomposed as

$$\int_{[1,e^{\hat{\theta}t}]}(\phi(\mathrm{id}, \hat{h}_-) - \phi(\mathrm{id}, h))\mathrm{d}h + \int_{[1,e^{\hat{\theta}t}]}\phi(\mathrm{id}, h)\mathrm{d}(\hat{h} - h)$$

$$+ \int_{[1,e^{\hat{\theta}t}]}(\phi(\mathrm{id}, \hat{h}_-) - \phi(\mathrm{id}, h))\mathrm{d}(\hat{h} - h) + O_P(n^{-1/2}). \tag{10.1}$$

All three terms can be bounded by $\|\hat{h} - h\|_\infty O_P(1)$. Therefore, the first assertion of the first theorem is an immediate consequence of Corollary 8.2.

The maximum likelihood estimator for G is obtained by applying the product integral to $\hat{\Lambda}_G$. Because the product integral is Lipschitz relative to the uniform norm if restricted to a domain of functions of uniformly bounded variation (see Gill 1994, Sections 2 and 4; or van der Vaart and Wellner 1996, p. 391), the uniform rate of $\hat{\Lambda}_G - \Lambda_G$ carries over onto the same uniform rate for $\hat{G} - G$. This proves the second assertion of Theorem 6.1.

By partial integration, the second term of (10.1) can be rewritten as

$$\phi(\mathrm{id}, h)(\hat{h} - h)\Big|_{1-}^{e^{\hat{\theta}t}+} - \int_{[1,e^{\hat{\theta}t}]}(\hat{h} - h)\mathrm{d}\phi(\mathrm{id}, h)$$

$$= \phi(\mathrm{id}, h)(\hat{h} - h)\Big|_{1-}^{e^{\hat{\theta}t}+} - \int_{[1,e^{\hat{\theta}t}]}\frac{\mathrm{d}}{\mathrm{d}s}\phi(s, h(s))\mathrm{d}(\hat{H} - H)(s). \tag{10.2}$$

The second term on the right of (10.2) is bounded by

$$2\|\hat{H} - H\|_\infty \int |\mathrm{d}^2/\mathrm{d}t^2\phi(t, h(t))|\mathrm{d}t = O_P(n^{-1/2}),$$

because we assume that $h''_{0,\theta}$ exists and is bounded. The first term on the right of (10.2), when multiplied by $n^{1/3}$, yields a non-trivial limit distribution by Theorem 8.5, which also shows that we may replace $e^{\hat{\theta}t}$ by $e^{\theta t}$.

Because the second-order partial derivative of ϕ relative to its second argument is bounded and $\int \mathrm{d}|h| + \mathrm{d}|\hat{h}| = O_P(1)$, the first and third terms of (10.1) change by at most $O_P(\|\hat{h} - h\|_\infty^2)$ if we replace the integrands by their linearization $\phi'_2(\mathrm{id}, h)(\hat{h}_- - h)$ in the second argument. Thus this change is $o_P(n^{-1/2})$ by Corollary 8.2. Next the linearization of the first term can be written

$$\int_{[1,e^{\hat\theta t}]} \phi_2'(\mathrm{id},\, h)h'\,\mathrm{d}(\hat H - H) = O_P(n^{-1/2}).$$

The linearization of the third term contributes

$$\int_{[1,e^{\hat\theta t}]} \phi_2'(\mathrm{id},\, h)(\hat h_- - h)\mathrm{d}(\hat h - h) = \tfrac{1}{2}\int_{[1,e^{\theta t}]} \phi_2'(\mathrm{id},\, h)\mathrm{d}(\hat h - h)^2 - \tfrac{1}{2}\sum_{[1,e^{\hat\theta t}]} \phi_2'(\mathrm{id},\, h)(\Delta\hat h)^2.$$

The absolute value of the first term on the right is bounded by

$$\|\hat h - h\|_\infty^2 \int |\mathrm{d}\phi_2'(\mathrm{id},\, h)| = o_P(n^{-1/2})$$

and hence is negligible at rate $n^{-1/3}$. However, the second term may contribute to the limit distribution of $(\hat\Lambda_G - \Lambda_G)(t)$. We conjecture that this term is $O_P(n^{-1/3})$, in which case the weak limit of the sequence $n^{1/3}(\hat\Lambda_G - \Lambda_G)(t)$ is the same as the weak limit of the sequence

$$\phi(t,\, h_{0,\theta}(t))n^{1/3}(\hat h_{0,\hat\theta} - h_{0,\theta})(e^{\theta t}) - \tfrac{1}{2}n^{1/3}\sum_{s\in[1,e^{\hat\theta t}]} \phi_2'(s,\, h_{0,\theta}(s))(\Delta\hat h_{0,\hat\theta})^2(s).$$

Here we note, by the characterization of $\hat h_{0,\theta}$ as the slope of $\hat H_{0,\theta}$, that $\hat h_{0,\hat\theta}(1-) - h_{0,\theta}(1) = \hat H_{0,\hat\theta}(1) - H_{0,\theta}(1) = o_P(n^{-1/3})$.

The weak limits of the sequences $n^{1/3}(\hat h_{0,\hat\theta} - h_{0,\theta})(e^{\theta t})$ can be shown to be asymptotically independent for different values of t. This suggests that the processes $t \mapsto n^{1/3}(\hat\Lambda_G - \Lambda_G)(t)$ will at best converge to weak limits in a pointwise sense and not uniformly in their argument t.

Next consider the proof of Theorem 6.2. For $t > 1 + a_n$ the continuous approximations $\tilde h_{0,\hat\theta}$ of $\hat h_{0,\hat\theta}$ satisfy

$$\tilde h_{0,\hat\theta}(t) - h_{0,\theta}(t) = \tfrac{1}{2}\int_{-1}^{1} (\hat h_{0,\hat\theta} - h_{0,\theta})(t - a_n u)\mathrm{d}u + O(a_n^2).$$

By this and a similar argument for the cumulative distribution functions, with $\|\cdot\|_{[a,b]}$ denoting the supremum norm on $[a, b]$,

$$\|\tilde h_{0,\hat\theta} - h_{0,\theta}\|_{[1+a_n,t-a_n]} \leqslant \|\hat h_{0,\hat\theta} - h_{0,\theta}\|_{[1,t]} + O(a_n^2),$$

$$\|\tilde H_{0,\hat\theta} - H_{0,\theta}\|_{[1+2a_n,t-a_n]} \leqslant \|\hat H_{0,\hat\theta} - H_{0,\theta}\|_{[1,t]} + O(a_n^2).$$

This implies that the preceding approximations remain valid if $\tilde h_{0,\hat\theta}$ is substituted for $\hat h_{0,\hat\theta}$, where now the term involving the jumps $\Delta\tilde h_{0,\hat\theta}$ vanishes, of course. Furthermore, for every fixed t,

$$\tilde h_{0,\hat\theta}(e^{\hat\theta t}) - \hat h_{0,\theta}(e^{\theta t}) = \tfrac{1}{2}\int_{-1}^{1} (\hat h_{0,\hat\theta}(e^{\hat\theta t} - a_n u) - \hat h_{0,\theta}(e^{\theta t}))\mathrm{d}u = o_P(n^{-1/3}),$$

in view of the second assertion of Theorem 8.5. The first assertion of Theorem 6.2 follows by Theorem 8.5. The second assertion follows by the delta method and the fact that, since $\tilde\Lambda_G$ is continuous, $1 - \tilde G(t) = e^{-\tilde\Lambda_G(t)}$.

References

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

Gill, R.D. (1989) Non- and semiparametric maximum likelihood estimators and the von-Mises method (part I). *Scand. J. Statist*., **16**, 97–128.

Gill, R.D. (1994) Lectures on survival analysis. In P. Bernard (ed.), *Lectures on Probability Theory: École d'Été de Probabilités de Saint-Flour XXII – 1992*, Lecture Notes in Math. 1581, pp. 115–241. New York: Springer-Verlag.

Gill, R.D. (1997) Nonparametric estimation under censoring and passive registration. *Statist. Neerlandica*, **51**, 35–54.

Groeneboom, P. (1985) Estimating a monotone density. In L. LeCam and R.A. Olshen (eds), *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. 2, pp. 539–555. Monterey, CA: Wadsworth.

Groeneboom, P. and Lopuhaa, H.P. (1993) Isotonic estimators of monotone densities and distribution functions. *Statist. Neerlandica*, **47**, 175–183.

Jonker, M.A. (2000) Statistical estimation of life length in historical demography. Doctoral thesis, Vrije Universiteit Amsterdam.

Murphy, S.A. and van der Vaart, A.W. (1999) Observed information in semi-parametric models. *Bernoulli*, **5**, 381–412.

Pfanzagl, J. with Wefelmeyer, W. (1982) *Contributions to a General Asymptotic Statistical Theory*, Lecture Notes in Statistics 13. New York: Springer-Verlag.

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

van der Vaart, A.W. (1996) Efficient estimation in semi-parametric models. *Ann. Statist*., **24**, 862–878.

van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist*., **20**, 595–601.

Wrigley, E.A. and Schofield, R.S. (1983) English population history from family reconstitution: Summary results. *Popul. Stud*., **37**, 157–184.

Wrigley, E.A., Davies, R.S., Oeppen, J.E. and Schofield, R.S. (1997) *English Population History from Family Reconstitution 1580–1837*. Cambridge: Cambridge University Press.

Received January 1999 and revised May 2000