Consistency of the NPML Estimator in the Right-Censored Transformation Model

E. V. SLUD

University of Maryland College Park

F. VONTA

University of Cyprus

ABSTRACT. This paper studies the representation and large-sample consistency for non-parametric maximum likelihood estimators (NPMLEs) of an unknown baseline continuous cumulative-hazard-type function and parameter of group survival difference, based on right-censored two-sample survival data with marginal survival function assumed to follow a transformation model, a slight generalization of the class of frailty survival regression models. The paper's main theoretical results are existence and unique a.s. limit, characterized variationally, for large data samples of the NPMLE of baseline nuisance function in an appropriately defined neighbourhood of the true function when the group difference parameter is fixed, leading to consistency of the NPMLE when the difference parameter is fixed at a consistent estimator of its true value. The joint NPMLE is also shown to be consistent. An algorithm for computing it numerically, based directly on likelihood equations in place of the expectation-maximization (EM) algorithm, is illustrated with real data.

Key words: adjoint differential equation, frailty model, large-sample theory, likelihood equation, restricted NPML, variational method

1. Introduction

A common problem in the analysis of clinical trials or epidemiological survival data is to infer the way in which survival over time depends upon auxiliary medical variables or risk-indicators, called *covariates*. Right-censored survival data are collected in the form of triples $(T_i, \Delta_i, Z_i) \in [0, \infty) \times \{0, 1\} \times \mathbf{R}^p$ for subjects i, and idealized in terms of the *latent failure model*, according to which each subject comes equipped with an unobserved random death-time X_i, random censoring time C_i, and discrete p-vector of covariates Z_i, with $T_i = \min(X_i, C_i)$ and $\Delta_i = I_{[X_i \leq C_i]}$. We impose the usual assumption that the vectors (X_i, C_i, Z_i) are independent and identically distributed, with C_i conditionally independent of X_i given Z_i. The objective is to estimate the conditional survival function $S(t|z)$ for X_i given Z_i.

By far the most common model for the influence of covariates is that of Cox (1972), according to which a factor depending upon covariates multiplies the hazard intensity. In this paper, we study a generalization of that model,

$$S(t|z) = \mathrm{P}\{X > t | Z = z\} = \exp(-G(e^{z'\beta}\Lambda(t))), \tag{1}$$

where G is assumed known and satisfies additional smoothness and regularity conditions discussed below. Both the true finite-dimensional coefficient-vector $\beta = \beta_0 \in \mathbf{R}^p$ and the true baseline continuous cumulative-hazard function Λ_0, are generally unknown. The problem of simultaneous estimation of (β, Λ) is called semiparametric because Λ is infinite-dimensional. Cox's (1972) model is the case $G(x) \equiv x$.

The groupwise survival functions $R_z(t) \equiv \mathrm{P}\{C > t | Z = z\}$ for censoring, as well as the laws of the random vectors $Z_i \in \mathbf{R}^p$, are assumed not to depend upon the parameters (β, Λ). In addition, all remaining study-subjects are right-censored at a fixed non-random time τ_0 such that $\Lambda(\tau_0) < \infty$. Equivalently, in terms of the group-z probabilities $c_z = P(Z_1 = z)$ and

$$q_z(t) \equiv P(T_1 \geq t, Z_1 = z) = c_z R_z(t) e^{-G(e^{z'\beta_0}\Lambda_0(t))}, \tag{2}$$

we have

$$\sum_z q_z(\tau_0) > 0 \quad \text{and} \quad \sum_z R_z(\tau_0+) = 0. \tag{3}$$

As a consequence, for all large n the longest durations in the observed dataset will almost surely be right-censored, with $\Delta_i = 0$, at time $T_i = \tau_0$. This is reasonable because biomedical studies will virtually never be continued until all subjects are dead. Even in accelerated-failure reliability studies, where all tested devices might be observed until failure, extremely delayed failures are more sensibly deemed censored, as accelerated stresses cannot be assumed to have the same effect on extremely long-lived devices as on others.

We study local maxima near (β_0, Λ_0) of the log-likelihood for model (1) of survival data $\{(T_i, \Delta_i, Z_i)\}_{i=1}^n$. Throughout most of the paper, $\rho \equiv e^\beta$ is a known positive scalar, with $Z_i \in \{0, 1\}$, and Λ, Λ_0 lie in the space of cumulative-hazard-like functions defined by

$$\mathcal{S}_0 \equiv \{\Lambda \in D[0, \tau_0] : \Lambda(0) = 0, \quad \Lambda \text{ non-decreasing}\},$$

and D is the space of right-continuous real-valued functions with left limits.

Denote the log-likelihood at Λ for the two-sample right-censored survival data under model (1), with $\rho \equiv e^\beta$, by logLik(Λ, ρ). The standard likelihood for right-censored survival data, with continuous Λ absolutely continuous with respect to the fixed dominating measure, is

$$\prod_{i=1}^n \left\{ \rho^{Z_i} G'(\rho^{Z_i}\Lambda(T_i)) \frac{d\Lambda}{dv}(T_i) e^{-G(\rho^z\Lambda(T_i))} \right\}^{\Delta_i} \left\{ e^{-G(\rho^z\Lambda(T_i))} \right\}^{(1-\Delta_i)}.$$

When the data $\{(T_i, \Delta_i, Z_i)\}_{i=1}^n$ are summarized through counting processes

$$N_z(t) = \sum_{i=1}^n \Delta_i I_{[T_i \leq t, Z_i = z]}, \quad Y_z(t) = \sum_{i=1}^n I_{[T_i \geq t, Z_i = z]},$$

and $N(t) = \sum_z N_z(t)$, $Y(t) = \sum_z Y_z(t)$, the logarithm of the likelihood is

$$\text{logLik}(\Lambda, \rho) = \sum_z \left\{ \int \log\left(\rho^z G'(\rho^z\Lambda(t)) \frac{d\Lambda}{dv} \right) dN_z(t) + \int G(\rho^z\Lambda(t)) dY_z(t) \right\}. \tag{4}$$

As in Nielsen *et al.* (1992), we define log-likelihood by the same formula more generally for $\Lambda \in \mathcal{S}_0$ with v from now on defined to be the sum of Λ_0 and counting measure for the jumps of N. Although standard, this choice will be justified later (in section 2).

The objective of this paper is to prove large-sample (local) existence and consistency of the generalized non-parametric maximum likelihood estimator (NPMLE) of (ρ, Λ) for model (1) when ρ is fixed in a sufficiently small neighbourhood (not depending on n) of ρ_0. We are interested in consistency of NPMLEs for Λ in the sense of uniform convergence on the compact interval $[0, \tau_0]$, where the non-random point τ_0 is as in (3) above.

Various authors have studied estimation in this setting, beginning with Clayton & Cuzick (1986) and Hougaard (1986). Dabrowska & Doksum (1988) proposed but did not rigorously justify an estimation method for frailty models. Nielsen *et al.* (1992) devised an estimator specifically for the Clayton–Cuzick frailty model, by modifying the EM algorithm. Klein (1992) implemented this estimator on real data, and Murphy (1994, 1995) established its asymptotic properties (consistency and asymptotic distribution). Klaassen (1993) proved existence of a consistent and efficient estimator of β in the uncensored Clayton–Cuzick model.

Cheng *et al.* (1995) and Bagdonavicius & Nikulin (1997) established asymptotic properties for estimating equation-based estimators in general transformation models. Parner (1998) has shown the joint NPMLE for (β, Λ) in the right-censored Clayton–Cuzick model to be consistent, asymptotically Gaussian, and semiparametric efficient. Murphy & van der Vaart (2000) present a theory of semiparametric profile likelihoods which would apply to our log-likelihood maximized over Λ for fixed ρ, but we cannot verify the hypotheses of their theorems in our setting.

This paper is organized as follows. Section 2 relates model (1) to survival frailty models (section 2.1) and gives general regularity conditions; then section 2.2 discusses alternative extensions of log-likelihood, and section 2.3 establishes equations for the NPMLE. Section 3 proves asymptotic absolute continuity for NPMLE sequences of Λ for fixed ρ, leading to a variational characterization of the limits for such sequences. Consistency results are collected in theorem 3 and corollary 2. In section 4, the NPML equations of theorem 1 lead to a convenient algorithm for calculation of the NPMLE. A related algorithm for simultaneous NPML estimation of ρ and Λ is illustrated for a previously analysed dataset of Christensen *et al.* (1985) on a clinical trial concerning primary biliary cirrhosis (PBC). In section 5, we sketch extensions of these results to right-censored regression models and to models with additional nuisance parameters such as the constant in the Clayton–Cuzick model. A brief discussion concludes the paper. Key technical calculations throughout the paper are deferred to appendices. Longer, more standard, calculations can be found in the report of Slud & Vonta (2002).

2. Background and assumptions

2.1. Frailty and transformation models

How do functions like G in model (1) arise? Most of those considered in the survival-analysis literature (cf. Clayton & Cuzick, 1986; Hougaard, 1986) derive from proportional hazard models (or Lehmann-alternatives two-group models) with an unobserved multiplicative random effect called *frailty*. That is, suppose there are, in addition to survival-time variables T and group-indicators Z, unobserved positive random variables ξ, and

$$S(t|Z = z, \xi) \equiv \exp(-\xi e^{z'\beta}\Lambda(t)).$$

The distribution function F_ξ of ξ may either be known, or known except for a parameter, but in the latter case it is crucial for identifiability of (β, F_ξ, Λ) that there not be two permissible F_ξ functions differing only by a scale change. Then the stratumwise unconditional survival function becomes

$$S(t|z) \equiv S(t|Z = z) \;=\; \int_0^\infty \exp(-xe^{z'\beta}\Lambda(t))dF_\xi(x) \equiv \exp(-G(e^{z'\beta}\Lambda(t))),$$

where

$$G(y) \;=\; -\ln\left(\int_0^\infty e^{-xy}dF_\xi(x)\right). \tag{5}$$

Model (1) and (5) with Gamma-distributed frailty, i.e. with F_ξ a $\Gamma(1/c,1/c)$ d.f., is the 'semiparametric Pareto model' of Clayton & Cuzick (1986). We refer to this special case, in which $G(x) = \ln(1+cx)/c$ for fixed $c > 0$, as the *Clayton–Cuzick model*. Murphy (1994, 1995) proved consistency and efficiency of the simultaneous NPMLE of (c, Λ) in this model when β (assumed $= 0$) is known. When F_ξ is the d.f. of a positive-stable r.v., $G(x) \equiv x^\alpha$ with $0 < \alpha \le 1$ as in Hougaard (1986).

Recently the family of semiparametric *transformation models*, which have been studied intensively in the case of uncensored data (Bickel *et al.*, 1993), have been extended for use with right-censored data. These models are coextensive with model (1), as can be seen from the formula $g(S(t|z)) = h(t) + z'\beta$ (1.3 of Cheng *et al.*, 1995), where g is known and h unknown, through the correspondence $g(x) \equiv \log(G^{-1}(-\log x))$, $h(t) \equiv \log \Lambda(t)$.

Throughout this paper, we are concerned with models $S(t|z)$ in formula (1) defined in terms of a known function G, about which we assume that:

(G.1) G is three times continuously differentiable, strictly increasing, and concave on $(0, \infty)$, with $G(0) = 0$;

(G.2) $-xG''(x)/G'(x)$ is uniformly bounded on $(0, \infty)$;

(G.3) $\int \exp(-G(x)) \log(G'(x))G'(x)dx > -\infty$; and

(G.4) $G'(0) < \infty$.

Condition (G.1) is easy to verify, via simple properties of Laplace transforms, when G arises as in (5) from a frailty model, but (G.1) holds more generally. Condition (G.2) holds for all frailty models with either $\inf(\mathrm{supp}(dF_\xi)) > 0$ or $F_\xi(\xi) \geq a\xi^b$ for ξ near 0, for some positive constants a, b. It is used to ensure dominatedness of the integrand in the integral for the expected gradient of logLik, which we need in order to differentiate under the integral sign. Condition (G.3) is a specialized assumption to make Kullback–Leibler information integrals finite, which holds in the most commonly applied frailty models, the Clayton–Cuzick and inverse-Gaussian (Hougaard, 1986) and positive-stable, but not in all frailty models. Condition (G.4), which excludes the positive-stable case, is needed in the proof of proposition 1. However, the positive-stable frailty model can be analysed separately using standard theory for the Cox model under the re-parameterization $\tilde{\Lambda} = \Lambda^\alpha$, $\tilde{\rho} = \rho^\alpha$.

2.2. Likelihood definition

As indicated above in the introduction, the log-likelihood under model (1) takes the form (4) when Λ is continuous. However, the extension of log-likelihood to functional parameters Λ which are allowed to have jumps can be made in several different ways. The one most accepted in the literature is due to Nielsen *et al.* (1992) and follows a clear train of thought, as follows. The likelihood for the model of Cox (1972), the special case of (1) in which $G(x) \equiv x$, had been written by Johansen (1983) as

$$\exp\left(\sum_z \int \left\{\log\left(\rho^z \frac{d\Lambda}{dv}\right) dN_z(t) - \rho^z Y_z d\Lambda\right\}\right), \tag{6}$$

for v from now on defined equal to the sum of Λ_0 and counting measure for a fixed countable set of (possible) jumps of Λ. The jumps of Λ enter this likelihood only in the terms $(\Delta\Lambda(t))^{\Delta N_z(t)} \exp(-\rho^z Y_z(t)\Delta\Lambda(t))$, with the interpretation that the failures at t consist of independent Poisson$(\rho^z\Delta\Lambda(t))$ numbers $\Delta N_z(t)$ in groups $z = 0, 1$. The likelihood (6) – which coincides with the usual censored survival data likelihood for continuous Λ – also makes sense for cumulative hazard functions Λ with jumps, and has the ideal property, established by Johansen (1983), that the Cox (1972) partial likelihood is derivable as (6) maximized over cumulative hazard functions Λ for fixed ρ. Nielsen *et al.*'s (1992) likelihood, with logarithm (4), is the one which results in frailty models (1) and (5) by: (i) replacing Λ in the

contribution to (6) of individual i by $\xi_i\Lambda$ for a F_ξ-distributed frailty variable ξ_i; and (ii) integrating out the ξ_i variable in the resulting expression with respect to the measure $dF_\xi(\xi_i)$. Nielsen *et al*. (1992) argue and Gill (1992) proves that this likelihood is also the one which results directly by consideration of intensities for the observed-data filtration, and according to Slud (1992) it can also be viewed as the limiting product of conditional likelihoods of observed-data increments over sequences of finer and finer partitions of the time-axis by stopping-time sequences.

The reasoning which led Johansen to the likelihood extension (6) for the Cox model also leads directly, in our setting with general G which may not arise from a frailty model, to the likelihood extension

$$\exp\left(\sum_z \int \left\{\log\left(\frac{d(G\circ\rho^z\Lambda)}{dv}\right)dN_z(t) + G(\rho^z\Lambda)dY_z\right\}\right). \tag{7}$$

Jumps in Λ contribute terms $(\Delta(G\circ\rho^z\Lambda)(t))^{\Delta N_z(t)}\exp(-Y_z(t)\Delta(G\circ\rho^z\Lambda)(t))$ to this likelihood, corresponding to independent Poisson$(\Delta(G\circ\rho^z\Lambda)(t))$ distributed numbers of failures at t in groups $z = 0, 1$. The logic supporting the logLik extension (7) is no more or less compelling than that of Johansen (1983). We argue in this way only to confirm that meaningful semi-parametric likelihood extensions *are not unique*. Another extension was studied, with methods like those of this paper, in Vonta (1992) and Slud & Vonta (2002). It seems likely, and is true in the Cox (1972) model although we cannot yet prove it in general, that the NPMLEs obtained by these variant likelihood extensions are all asymptotically equivalent.

2.3. NPML equations

We first find necessary conditions for a maximum at $\vartheta = 0$ of log-likelihood under model (1) over one-parameter families $\Lambda = \Lambda_\vartheta$ defined (Gill 1989) by

$$\Lambda_\vartheta(t) \equiv \int_0^t (1 + \vartheta\cdot\gamma(s))d\Lambda(s),$$

where ϑ varies over a small neighbourhood of 0, and γ is a bounded measurable function on $[0, \tau_0]$. This approach leads to *local* or *relative* NPMLEs in the sense of Kiefer & Wolfowitz (1956). The maxima are often taken for fixed values ρ (generally different from ρ_0), in which case we speak of *restricted NPMLEs*.

As we employ the same logLik extension (4) as Nielsen *et al*. (1992) and virtually all later authors, the joint NPMLEs we study agree precisely with the Cox (1972) maximum partial likelihood estimators in the case $G(x) = x$, according to Johansen (1983), and to the NPMLEs studied by Klein (1992), Murphy (1994, 1995) and Parner (1998) in the Clayton–Cuzick model. However, the asymptotic behaviour of restricted NPMLEs has not previously been studied in models (1) other than Cox's.

The space $\mathcal{S}_0$ of allowable functions Λ is large enough to contain all potential NPMLEs. It suffices to check that an extended real-valued Λ which attains the value ∞ within $[0, \tau_0]$ already makes logLik negatively infinite. To show this, we first take v to be the sum of $d\Lambda_0$ and a counting measure on a countable set of potential jumps, and rewrite (4) as

$$\sum_z\left\{\int \log\left(\rho^z\left(\frac{d\Lambda}{d\Lambda_0}+\Delta\Lambda\right)\left(\frac{G'}{G}\right)\Big|_{\rho^z\Lambda}(Ge^{-G})\Big|_{\rho^z\Lambda}\right)dN_z + \int G(\rho^z\Lambda)d(Y_z+N_z)\right\}. \tag{8}$$

Now remark that $Y_z + N_z$ is non-increasing on $[0, \tau_0]$ by definition; that $Y(\tau_0) > 0$ while $Y(\tau_0+) = 0$ by (3), and that $yG'(y)/G(y) \le 1$ for $y \in (0, \infty)$ by concavity of G. Therefore the

second integral in (8) is at most $-G(\rho^z\Lambda(\tau_0))Y_z(\tau_0)$, and the argument of the logarithm of the first integral is finite at all points of $(0, \tau_0)$, including any t for which $\Lambda(t) = \infty > \Lambda(t-)$. This shows

Lemma 1
logLik(Λ, ρ) *as displayed in (8) is only increased, as a function on the space of extended real-valued cumulative hazard functions on* $[0, \tau_0]$, *if* Λ *is restricted to lie in* $\mathcal{S}_0$, *i.e. to satisfy* $\Lambda(\tau_0) < \infty$.

The first Gâteaux derivative $(\partial/\partial\vartheta)\,\mathrm{logLik}(\Lambda_\vartheta, \rho)|_{\vartheta=0}$ at the parameter-location Λ in the direction $\Lambda_\vartheta - \Lambda$ has the general formula

$$\sum_z \Bigg\{ \int \left(\gamma(t) + \rho^z \frac{G''(\rho^z\Lambda(t))}{G'(\rho^z\Lambda(t))} \left(\int_0^t \gamma d\Lambda \right) \right) (I_{[\Delta\Lambda(t)=0]} + I_{[\Delta\Lambda(t)>0]}) dN_z(t) + \int (-Y_z(t))\rho^z I_{[\Delta\Lambda(t)>0]} \left\{ \left(\int_0^{t-} \gamma d\Lambda \right) d(G' \circ \rho^z\Lambda)(t) + G'(\rho^z\Lambda(t))\gamma(t) d\Lambda \right\} + \int (-Y_z(t))\rho^z \left\{ G''(\rho^z\Lambda(t))\rho^z \left(\int_0^t \gamma d\Lambda \right) + G'(\rho^z\Lambda(t))\gamma(t) \right\} I_{[\Delta\Lambda(t)=0]} d\Lambda \Bigg\}, \tag{9}$$

for all $\Lambda \in \mathcal{S}_0$ and all bounded measurable γ. Lemmas 3 and 4 of appendix A show that a necessary condition for (9) to be 0 for all bounded measurable γ, for fixed (Λ, ρ), is that Λ be a pure-jump function with jumps occurring at precisely the locations of jumps of N, i.e. for

$$\mathcal{S}_n \equiv \{\Lambda \in \mathcal{S}_0 : \text{for } t \geq 0,\ I_{[\Delta\Lambda=0]} dN + I_{[\Delta N=0]} d\Lambda \equiv 0\}, \tag{10}$$

$$\Lambda \quad \text{is a NPMLE} \quad \Longrightarrow \quad \Lambda \in \mathcal{S}_n. \tag{11}$$

The non-zero terms in formula (9) for such Λ result in a tractable finite set of NPML equations, stated here and proved in appendix A.

Theorem 1
For $\Lambda \in \mathcal{S}_n$ *to be an NPMLE for model (1) based on the data* $(N_z(t), Y_z(t), z = 0, 1)_{t\geq 0}$ *with fixed* $\rho = e^\beta$, *the following system of equations must hold: if* $s < t$ *are any two successive jumps of N, and* t_* *is the last jump of N, then*

$$\sum_z \rho^z \left\{ G'(x)(Y_z(s) - Y_z(t)) - \frac{G''(x)}{G'(x)} \Delta N_z(s) \right\}_{x=\rho^z\hat\Lambda(s)} = \frac{\Delta N(s)}{\Delta\hat\Lambda(s)} - \frac{\Delta N(t)}{\Delta\hat\Lambda(t)}, \tag{12}$$

$$\sum_z \rho^z \left\{ (-Y_z(t_*))G'(\rho^z\hat\Lambda(t_*)) + \frac{G''}{G'}\Big|_{\rho^z\hat\Lambda(t_*)} \Delta N_z(t_*) \right\} + \frac{\Delta N(t_*)}{\Delta\hat\Lambda(t_*)} = 0. \tag{13}$$

3. Existence and consistency of NPMLE

By (11), $\Lambda \in \mathcal{S}_n$ is necessary for Λ to be an NPMLE. When $\Lambda \in \mathcal{S}_n$, (4) provides $\mathrm{logLik}(\Lambda, \rho)$ equal to

$$\sum_z \int (G \circ \rho^z\Lambda) dY_z + \sum_{z,t} \log(\rho^z G'(\rho^z\Lambda(t))\Delta\Lambda(t))\Delta N_z(t). \tag{14}$$

We study the maximization of logLik for fixed ρ by grouping the terms in the last summation of (14) in a special way. Fix, for all large n, a non-random finite system

$\underline{\gamma} = \{\gamma_{in}\}_{i=1}^{m(n)} = \{\gamma_i\}_{i=1}^{m}$ of intervals $(\gamma_i, \gamma_{i+1}]$ partitioning $(0, \tau_0]$, where $0 = \gamma_0 < \gamma_1 < \cdots < \gamma_m < \gamma_{m+1} = \tau_0$, and satisfying the following condition, with $\lfloor \cdot \rfloor$ denoting greatest integer:

(P.1) $\Lambda_0(\gamma_{i+1}) - \Lambda_0(\gamma_i) = \frac{\Lambda_0(\tau_0)}{m}, \quad i \le m = m(n),$
where $m(n) = \lfloor \Lambda_0(\tau_0)\sqrt{n} \cdot \min_z q_z(\tau_0)\rho_0^z G'(\rho_0^z \Lambda_0(\tau_0)) \rfloor$.

As the underlying distributions of the failure-times X_i are continuous, a.s. $\Delta N(\gamma_i) = 0$ for all i. Now define for each $i \le m$, $z = 0, 1$,

$$r_{i,z} = N_z(\gamma_{i+1}) - N_z(\gamma_i) = \sum_{t \in (\gamma_i, \gamma_{i+1}]} \Delta N_z(t),$$

$$\pi_{i,z} = E\left(\frac{r_{i,z}}{n}\right) = \int_{\gamma_i}^{\gamma_{i+1}} q_z(t) d(G \circ \rho_0^z \Lambda_0)(t),$$

$$C_{i,z}(\Lambda) = \sum_{t \in (\gamma_i, \gamma_{i+1}]} [\rho^z G'(\rho^z \Lambda(t)) \Delta\Lambda(t)] \Delta N_z(t), \quad \Lambda \in \mathcal{S}_0.$$

Note that all of the quantities γ_i, m, r_{iz}, π_{iz}, $C_{iz}(\Lambda)$ depend upon n, but for convenience we suppress this dependence.

In terms of these notations, we have

$$\log\mathrm{Lik}(\Lambda, \rho) = \sum_z \int (G \circ \rho^z \Lambda) dY_z + \sum_{j,z} r_{j,z} \left\{ \log(C_{j,z}(\Lambda)) + \sum_{t: t \in (\gamma_j, \gamma_{j+1})} \left(\frac{\Delta N_z(t)}{r_{j,z}}\right) \log\left(\frac{\rho^z G'(\rho^z \Lambda(t)) \Delta\Lambda(t)}{C_{j,z}(\Lambda)}\right) \right\}. \tag{15}$$

The idea of grouping terms in just this way is that for fixed (j, z), each of the vectors of dimension $r_{j,z}$ indexed by the jump-points t for N_z within $(\gamma_j, \gamma_{j+1}]$ with components

$$\frac{1}{r_{j,z}} \quad \text{and} \quad \frac{\{\rho^z G'(\rho^z \Lambda(t)) \Delta\Lambda(t)\}}{C_{j,z}(\Lambda)}$$

is a *probability vector*. It is a simple consequence of Jensen's inequality that for a fixed positive probability vector $\mathbf{x}$ of finite dimension d,

$$\max\left\{ \sum_k x_k \log p_k : \quad \mathbf{p} \in [0, 1]^d, \sum_k p_k = 1 \right\} = \sum_k x_k \log x_k. \tag{16}$$

So we have proved an upper bound for logLik in the following lemma which will turn out to be asymptotically attainable.

Lemma 2
For arbitrary $\Lambda \in \mathcal{S}_n$, *if* $N_z(\gamma_j) < N_z(\gamma_{j+1})$ *for all* $j \le m$ *and* $z = 0, 1$,

$$\mathit{logLik}(\Lambda, \rho) \le \sum_z \int (G \circ \rho^z \Lambda) dY_z + \sum_z \sum_j r_{j,z} \log\left(\frac{C_{j,z}(\Lambda)}{r_{j,z}}\right).$$

The log-likelihood at (Λ, ρ) *is negatively infinite if either* $\Lambda(\gamma_j) = \infty$ *or* $\Lambda(\gamma_j) = \Lambda(\gamma_{j+1})$ *for some* $j \le m$.

From condition (P.1) along with properties (G.1) and (G.4) of G, the following useful properties of $\{\gamma_i, r_{i,z}: 0 \le i \le m + 1, z = 0, 1\}$ are derived in Slud & Vonta (2002): for $z = 0, 1$, a.s. for all sufficiently large n,

$$\text{for all} \quad i \le m, \ a.s. \quad \pi_{i,z} \ge 1/\sqrt{n} \quad \text{and} \quad |r_{i,z} - n\pi_{i,z}| \le n^{3/8}, \tag{17}$$

$$\max_{z=0,1, i\le m} \left\{ \left| \log\left(\frac{G'(\rho^z \Lambda_0(\gamma_{i+1}))}{G'(\rho^z \Lambda_0(\gamma_i))} \right) \right| + \left| \log\left(\frac{q_z(\gamma_i)\rho^z G'(\rho^z \Lambda_0(\gamma_i))(\Lambda_0(\gamma_{i+1}) - \Lambda_0(\gamma_i))}{\pi_{i,z}} \right) \right| \right\} \to 0. \tag{18}$$

3.1. Limiting behaviour of NPMLEs

The maximizers of log-likelihood over $\Lambda \in \mathcal{S}_0$ have been shown in (11) to lie in the finite-dimensional set $\mathcal{S}_n$, and for fixed ρ, logLik(Λ, ρ) is continuous in Λ and tends to $-\infty$ as Λ tends to the boundary of $\mathcal{S}_n$. Therefore, the maximization can without loss of generality also be restricted to a compact subset of $\mathcal{S}_n$, and it follows immediately that relative maximizers $\hat{\Lambda} \in \mathcal{S}_n$ of the log-likelihood (4) do exist. In the next two results, proved in appendix B, we establish further properties which must almost surely be satisfied by any NPMLE sequence Λ_n based on data samples of size n as $n \to \infty$.

Proposition 1

Assume model (1) *with fixed underlying parameters* (Λ_0, ρ_0) *and continuous* Λ_0, *for survival-data samples* $\{(T_i, \Delta_i, Z_i), i = 1, \ldots, n\}$. *Let* ρ *be fixed* (*not necessarily equal to* ρ_0), *and let* $\{\gamma_j\}_{j=1}^m$ *satisfy (P.1) and* (18) *as above. Then for any finite constant* $K > \Lambda_0(\tau_0)$, *there exists a finite constant C not depending upon n, such that if* $\{\Lambda_n \in \mathcal{S}_n\}$ *is any sequence of relative maximizers of (4) within* $\{\Lambda \in \mathcal{S}_n : \Lambda(\tau_0) \le K\}$, *then almost surely for all sufficiently large n, for all* $j \le m$,

$$\Lambda_n(\gamma_{j+1}) - \Lambda_n(\gamma_j) \le C \cdot (\Lambda_0(\gamma_{j+1}) - \Lambda_0(\gamma_j)). \tag{19}$$

Theorem 2

Under the hypotheses of proposition 1,

$$\limsup_n \frac{1}{n} [\text{logLik}(\Lambda_n) + N(\infty) \log n] \le \sup_L \mathcal{J}(L, \rho), \tag{20}$$

where the supremum is taken over non-decreasing functions L absolutely continuous with respect to Λ_0 (*i.e. such that the corresponding measures satisfy* $dL \ll d\Lambda_0$), *and the objective-functional* $\mathcal{J}(L, \rho)$ *is defined for functions* $L \in \mathcal{S}_0$ *by*

$$\mathcal{J}(L, \rho) = \sum_z \left[\int (G \circ \rho^z L) dq_z + \int q_z \log \frac{d(G \circ \rho^z L)}{d\Lambda_0} d(G \circ \rho_0^z \Lambda_0) \right] - \sum_z \int q_z \log \left(\sum_{w=0}^{1} q_w \rho_0^w G'(\rho_0^w \Lambda_0) \right) d(G \circ \rho_0^z \Lambda_0). \tag{21}$$

Corollary 1

Under the same hypotheses as proposition 1, if an NPMLE sequence $\{\Lambda_{n'}\}$ *along a subsequence of samples of size* n' *falls within a set*

$$A \equiv \{\Lambda \in \mathcal{S}_0 : \sup_{t \le \tau_0} |\Lambda(t) - L_0(t)| \ge \delta\},$$

for a fixed function $L_0 \in \mathcal{S}_0$ *and* $\delta > 0$ *not depending upon* n', *then*

$$\limsup_{n'} \frac{1}{n'} (logLik(\Lambda_{n'}, \rho) + N(\infty) \log n') \le \sup_{L \in A} \mathcal{J}(L, \rho),$$

where the supremum runs without loss of generality over functions $L \in \mathcal{S}_0$ *absolutely continuous with respect to* Λ_0, *as in the theorem.*

3.2. Variational characterization of NPMLE limits

We next study the variational problem of maximizing $\mathcal{J}(\cdot,\rho)$ over elements $L \in \mathcal{S}_0$ which are absolutely continuous with respect to Λ_0. The expression $\mathcal{J}(L, \rho)$, without the last line of (21) which is free of L, is the expectation under the true parameter values (ρ_0, Λ_0) of n^{-1} logLik(L, ρ) for data right-censored no later than τ_0. It is verified in Slud & Vonta (2002) that the functions Λ for which $\mathcal{J}(L, \rho) > -\infty$ all satisfy

$$L \ll \Lambda_0 \quad \text{and} \quad \sum_z \int q_z G'(\rho^z L) dL < \infty. \tag{22}$$

Now fix a parameter value (ρ, L) satisfying (22), with $L(\tau_0) < \infty$, and consider the one-parameter family of functions $L_\vartheta(u) = \int_0^u \psi(\vartheta h(s))\, dL(s)$ in the neighbourhood of $\vartheta = 0$, where ψ is an arbitrary fixed smooth non-decreasing scalar function from the whole real line to the positive half-line such that $\psi(0) = 1$, $\psi'(0) = 1$, and where h is an arbitrary element of the linear space

$$H_{\delta,L} = \left\{ h \in L_0(\mathbf{R}^+, d\Lambda_0) : \sup_{z,\rho \in B_\delta(\rho_0)} \int_0^{\tau_0} h^2 R_z dL < \infty \right\}.$$

We introduce ψ, following Bickel *et al.* (1993), in order that the directions h fill out a linear space, without awkward constraints. The (small) constant δ will be chosen below.

The condition that $E(\text{logLik}(L_\vartheta, \rho))$ is extremized with respect to ϑ at 0 for arbitrary h, i.e. has horizontal derivative, can be written

$$\sum_z \int \left(h + \rho^z \frac{G''}{G'}\Big|_{\rho^z L} \int_0^{\cdot} h dL \right) q_z \left\{ \rho_0^z G'(\rho_0^z \Lambda_0) d\Lambda_0 - \rho^z G'(\rho^z L) dL \right\} = 0.$$

After integration by parts, we obtain as an equivalent condition for this extremum at L to hold simultaneously for all h, that for $s < \tau_0$:

$$\sum_z \left[\rho_0^z G'(\rho_0^z \Lambda_0(s)) q_z(s) - \frac{dL}{d\Lambda_0} \rho^z \left(q_z(s) G'(\rho^z L(s)) + \rho^z \int_s^{\tau_0} G''(\rho^z L) q_z dL(u) \right.\right.$$
$$\left.\left. - \rho_0^z \int_s^{\tau_0} \frac{G''}{G'}\Big|_{\rho^z L} G'(\rho_0^z \Lambda_0(u)) q_z d\Lambda_0(u) \right) \right] = 0,$$

or

$$\frac{dL}{d\Lambda_0}(s) = \frac{\sum_z \rho_0^z q_z(s) G'(\rho_0^z \Lambda_0(s))}{\sum_z \rho^z \left[q_z(s) G'(\rho^z L(s)) - \int_s^{\tau_0} q_z (G''(\rho^z L)/G'(\rho^z L)) d\left(G \circ \rho_0^z \Lambda_0 - G \circ \rho^z L \right) \right]}. \tag{23}$$

Next re-parameterize (23) using

$$\alpha = -\sum_z \rho^z \int_0^{\tau_0} q_z \frac{G''}{G'}\Big|_{\rho^z L} (d(G \circ \rho_0^z \Lambda_0) - d(G \circ \rho^z L)), \tag{24}$$

so that the right-hand side of equation (23) becomes

$$\frac{\sum_z \rho_0^z G'(\rho_0^z \Lambda_0(s)) q_z(s)}{\alpha + \sum_z \rho^z \left[q_z(s) G'(\rho^z L(s)) + \int_0^s q_z (G''/G')\Big|_{\rho^z L} (d(G \circ \rho_0^z \Lambda_0) - d(G \circ \rho^z L)) \right]}.$$

By (G.1) and (G.4) making G''/G' bounded on a neigbourhood $[0, \epsilon]$ of 0, and by (G.2) making it bounded on each interval $[\epsilon, \tau_0]$, this parameter α will be finite for any L satisfying (22). Now treating α as an unknown parameter and defining

$$P(s) = P(s, \alpha, \rho) = \alpha + \sum_z \rho^z \int_0^s q_z \frac{G''}{G'}\Big|_{\rho^z L} (d(G \circ \rho_0^z \Lambda_0) - d(G \circ \rho^z L)),$$

we transform equation (23) to obtain the second order system

$$\frac{dL}{d\Lambda_0}(s) = \frac{\sum_z \rho_0^z q_z(s) G'(\rho_0^z \Lambda_0(s))}{\sum_z \rho^z q_z(s) G'(\rho^z L(s)) + P(s)} \tag{25}$$

$$\frac{dP}{d\Lambda_0}(s) = \sum_z \rho^z q_z(s) \frac{G''}{G'}\Big|_{\rho^z L} \left(\rho_0^z G'(\rho_0^z \Lambda_0(s)) - \rho^z G'(\rho^z L(s)) \frac{dL}{d\Lambda_0}(s) \right) \tag{26}$$

$$L(0) = 0, \quad P(0) = \alpha. \tag{27}$$

It follows from the particular choice of α defined in (24), for a calculus-extremum L of $\mathcal{J}(\cdot, \rho)$, that $P(\tau_0) = 0$, and we seek to characterize α in this way implicitly but uniquely from the system (25)–(26).

Proposition 2

For all ρ lying in a sufficiently small interval $(\rho_0 - \delta, \rho_0 + \delta)$, the expression $\mathcal{J}(L, \rho)$ in (21) *is uniquely maximized over non-decreasing $L \in \mathcal{S}_0$ at the function $L = L_\rho$ which solves the equation system* (25)–(26) *subject to the conditions $L(0) = 0$, $P(\tau_0) = 0$.*

Proof. A solution (L, P) of (25)–(27) for $\rho = \rho_0$, $\alpha = 0$ is given on $[0, \tau_0]$ by $L(s) \equiv \Lambda_0(s)$, $P(s) \equiv 0$, and the right-hand sides of (25)–(26) are uniformly bounded and smooth on $[0, \tau_0]$ and smooth for (α, ρ) in a neighbourhood of $(0, \rho_0)$. Standard ordinary differential equation (ODE) theory (cf. Coddington & Levinson 1957, Chapter 1) and the smoothness assumptions on G imply that for (α, ρ) in a sufficiently small neighbourhood $\mathcal{U}$ of $(0, \rho_0)$, the solution (L, P) of (25)–(27) on $[0, \tau_0]$ depends smoothly on the parameters (α, ρ). Thus, there exists a rectangular neighbourhood $(-\delta, \delta) \times (\rho_0 - \delta, \rho_0 + \delta) \subset \mathcal{U}$ of values (α, ρ) such that the solutions $(L(s, \alpha, \rho), P(s, \alpha, \rho))$ are continuously differentiable with respect to s, α. We study next the behaviour of the partial derivatives

$$L_*(s) \equiv \frac{\partial L}{\partial \alpha}(s, 0, \rho_0), \quad P_*(s) \equiv \frac{\partial P}{\partial \alpha}(s, 0, \rho_0), \tag{28}$$

when (α, ρ) lies in a small neighbourhood of $(0, \rho_0)$, by means of the adjoint system obtained by formal differentiation of the system (25)–(26) with respect to α at the point $(\alpha, \rho) = (0, \rho_0)$:

$$\frac{dL_*}{d\Lambda_0}(s) = -\frac{P_*(s) + \sum_z \rho_0^{2z} G''(\rho_0^z \Lambda_0(s)) q_z(s) L_*(s)}{\sum_z \rho_0^z G'(\rho_0^z \Lambda_0(s)) q_z(s)} \tag{29}$$

$$\frac{dP_*}{d\Lambda_0}(s) = -\sum_z \rho_0^z q_z(s) \left(\frac{G''}{G'}\right)_{\rho_0^z \Lambda_0} \left(\rho_0^{2z} G''(\rho_0^z \Lambda_0) L_* - \frac{\rho_0^z G'(\rho_0^z \Lambda_0)(P_* + \sum_w \rho_0^{2w} G''(\rho_0^w \Lambda_0) q_z L_*)}{\sum_w \rho_0^w q_w G'(\rho_0^w \Lambda_0)} \right), \tag{30}$$

subject to the initial condition $L_*(0) = 0$, $P_*(0) = 1$. Under our regularity assumptions, the solution of (29)–(30) exists and is unique, and satisfies

$$\inf_{s \in [0,\tau_0]} P_*(s) > 0, \tag{31}$$

which is proved in appendix C. Therefore $(\partial P/\partial\alpha)(\tau_0, 0, \rho_0) > 0$, so that, possibly after making the neighbourhood still smaller, for some positive constant b, and neighbourhood $\mathcal{U}_1$ of $(0,\rho_0) \in \mathbf{R} \times \mathbf{R}^+$,

$$L(\tau_0, \alpha, \rho) \geq b, \quad \frac{\partial P}{\partial \alpha}(\tau_0, \alpha, \rho) \geq b, \quad |P(\tau_0, \alpha, \rho)| \leq \frac{b\tau_0}{2}. \tag{32}$$

Thus, for each ρ close enough to ρ_0, there is a unique $\alpha = \alpha(\rho)$ such that

$$(\alpha(\rho), \rho) \in \mathcal{U}_1 \quad \text{and} \quad P(\tau_0, \alpha(\rho), \rho) = 0.$$

Moreover, the inverse function theorem implies that this locally defined function $\alpha(\rho)$ is continuously differentiable.

The reasoning immediately preceding the statement of the proposition showed that a local extremum of the functional $\mathcal{J}(\cdot, \rho)$ on $\mathcal{S}_0$ for fixed ρ must necessarily satisfy (25)–(26) with a real parameter α and function P such that $P(0) = \alpha$, $P(\tau_0) = 0$. For $\rho \in (\rho_0 - \delta, \rho_0 + \delta)$, such solutions exist *and are unique* within the set of $\mathcal{S}_0$ functions which are bounded and have bounded density derivatives with respect to Λ_0 on $[0, \tau_0]$. However, the collection of such continuous functions on $[0, \tau_0]$ is relatively compact in uniform norm by the Arzela–Ascoli theorem. Hence the continuous functional $\mathcal{J}(\cdot, \rho)$ on $(\mathcal{S}_0, \|\cdot\|_{\infty,[0,\tau_0]})$ has a maximizer. In summary, the functions $L_\rho \equiv L(\cdot, \alpha(\rho), \rho)$ smoothly parameterized by ρ are each unique local maximizers of $\mathcal{J}(\cdot, \rho)$ over $\mathcal{S}_0$, as was to be proved.

The function L_ρ characterized in proposition 2 is the unique function to which the NPMLE based on the fixed local value ρ a.s. converges.

3.3. Consistency theorems

Theorem 3

Denote by $\rho_0 = e^{\beta_0}$ and Λ_0 the true values of the parameters ρ and Λ governing the data $\{N_z(t), Y_z(t), z = 0,1, t \geq 0\}$ under model (1)*, together with* (3) *and* (G.1)–(G.4)*; and assume that the groupwise sample-sizes $n_z \equiv Y_z(0)$ grow with n in such a way that almost surely*

$$n_z/n \to c_z \text{ as } n \to \infty, \tag{33}$$

where $c_z > 0$ are constants. Then almost surely, for each fixed ρ in a sufficiently small interval $(\rho_0 - \delta, \rho_0 + \delta)$ and each sufficiently large sample size n, there exists an NPMLE sequence Λ_n which satisfies $\limsup_{n \to \infty} \Lambda_n(\tau_0) < \infty$. For every such NPMLE sequence,

$$\lim_n \frac{1}{n}[logLik(\Lambda_n) + N(\infty)\log n] = \mathcal{J}(L_\rho, \rho), \tag{34}$$

where $\mathcal{J}(\cdot,\rho)$ is the functional (21)*, and where the unique solution L_ρ of the differential equations* (25)–(26) (*with auxiliary function $P(\cdot)$ such that $P(\tau_0) = 0$*) *is also the unique maximizer of $\mathcal{J}(\cdot,\rho)$. Moreover,*

$$\sup_{t \in [0,\tau_0]} |\Lambda_n(t) - L_\rho(t)| \longrightarrow 0 \quad \text{as} \quad n \to \infty. \tag{35}$$

The proof of this theorem is given in appendix D. This proof shows that when ρ is fixed at ρ_0, the functions L_ρ determined by the system (25)–(26) of ODEs converge uniformly on compact sets as $\rho \to \rho_0$ to $L_{\rho_0} = \Lambda_0$. Therefore, we have also proved:

Corollary 2
Under the assumptions of theorem 3, if ρ is fixed either precisely at ρ_0 or at a strongly consistent estimator $\tilde{\rho}_n$ of ρ_0, then there exist NPMLEs with values at τ_0 bounded for all large n by $\Lambda_0(\tau_0) + 1$, and any such sequence $\hat{\Lambda}_n$ of NPMLEs is strongly consistent as $n \to \infty$.

Theorem 3 and corollary 2 imply that any (restricted) NPMLE sequence over the set $\mathcal{S}_n \cap \{\Lambda : \Lambda(\tau_0) \le \Lambda_0(\tau_0) + 1\}$ will be consistent, a remark with a very attractive computational consequence.

Corollary 3
Under the assumptions of theorem 3, define $\hat{\Lambda}_{KMz}$ to be the Nelson–Aalen cumulative hazard estimator based on the data from group $z = 0, 1$, and define the estimator $\tilde{\rho}$ of ρ_0 by

$$\tilde{\rho} = \frac{G^{-1}(\log 2)}{\hat{\Lambda}_{KM1}(\hat{\Lambda}^{-1}_{KM0}(G^{-1}(\log 2)))},$$

in terms of right-continuous inverses. Then any restricted NPMLE of Λ_0 within $\mathcal{S}_n \cap \{\Lambda : \Lambda(\tau_0) \le \hat{\Lambda}_{KM0}(\tau_0) + 1\}$, with ρ fixed either at or in a small neighbourhood of $\tilde{\rho}$, is a consistent estimator, respectively, of Λ_0 or L_ρ.

The idea of the preliminary estimators used in corollary 3 is first that the groupwise Kaplan–Meier estimators are consistent, second that the (smallest Λ_0 support-point greater than or equal to the) corresponding group-0 median survival time is consistent, and therefore that $\tilde{\rho}$ consistently estimates ρ_0.

4. Numerical algorithm and data example

4.1. Algorithm for estimation

Let $t_{(i)}$, $i = 1, \ldots, r$ denote the ordered jump-times for N, and $z_{(i)}$ denote the corresponding group-indicators of the individuals failing at these times, where $r = N(\tau_0)$. Equation (12) for the NPMLE, given in theorem 1, says for $i = 1, \ldots, r-1$ that for fixed ρ

$$\frac{\rho^{z_{(i)}} G''(\rho^{z_{(i)}} \Lambda(t_{(i)}))}{G'(\rho^{z_{(i)}} \Lambda(t_{(i)}))} + \frac{1}{\Delta\Lambda(t_{(i)})} - \frac{1}{\Delta\Lambda(t_{(i+1)})} = \sum_z \rho^z G'(\rho^z \Lambda(t_{(i)}))(Y_z(t_{(i)}) - Y_z(t_{(i+1)})), \tag{36}$$

while (13) says that

$$\Lambda(t_{(r-1)}) = \Lambda(t_{(r)}) - \frac{1}{\sum_z \rho^z Y_z(t_{(r)}) G'(\rho^z \Lambda(t_{(r)})) - (\rho^{z_{(r)}} G''(\rho^{z_{(r)}} \Lambda(t_{(r)}))/G'(\rho^{z_{(r)}} \Lambda(t_{(r)})))}. \tag{37}$$

Equations (36) and (37) enable a backwards induction according to which $\hat{\Lambda}(t_{(k-1)})$ is determined uniquely from $\hat{\Lambda}(t_{(i)})$, $i \ge k$ for k ranging from r to 1. In these recursions, $\Lambda(t_r)$ is a finite unknown constant, to be determined from the system of equations (36) and (37),

along with equation (36) at $i = 0$, where $t_{(0)} = 0$ by definition, and $\Lambda(t_{(0)})$ must be set equal to 0.

Theorem 1 and the backward recursions (36) and (37) lead immediately to a strategy for constructing the restricted NPMLE $\hat{\Lambda} = \hat{\Lambda}_\rho$ for fixed ρ. The idea is analogous to the 'shooting method' in numerically solving two-point boundary-value differential-equation problems. For fixed ρ, define the function $D(u)$ of the starting value $u = \Lambda(t_{(r)}) = \Lambda(\tau_0)$ for the recursion, to be equal to the value of $\Lambda(t_{(0)})$ obtained by following the recursion (36) or (37) back to $i = 0$. Then the estimated value $\hat{\Lambda}(t_{(r)})$ is defined as the value for which $D(u) = 0$. In the present setting, this function $D(\cdot)$ is well-defined, and the root of $D(u) = 0$ has always been found in our numerically computed examples, although we are not able to prove it generally, to be unique. For each root $u = \Lambda(t_{(r)})$ of $D(u) = 0$, there will exist a corresponding value of $\Lambda(t_{(1)})$, and we remark that the entire sequence of values $\Lambda(t_{(k)})$, $k \geq 2$, can be recovered from $\Lambda(t_{(1)})$ via a forward recursion using (36) and (37).

Suppose we have fixed $\Lambda(t_{(j)})$, $j = 0, 1, \ldots, i$, $i \geq 1$. Then there is only one possible value of $u = \Lambda(t_{(i+1)})$ (the root of a monotone decreasing function of u) as can be observed by (36). This reasoning shows that the roots $u = \Lambda(t_{(r)})$ of D stand in one-to-one correspondence with the associated values $\Lambda(t_{(1)})$. As a root-finder for $D(\cdot)$ requires a starting point, it is reasonable to begin by using a simple consistent empirical estimator $\tilde{L}_\rho$ for the limit L_ρ of NPMLEs for fixed ρ. The choice of the $\tilde{L}_\rho$ estimator was given in corollary 3. The suggested initial value for u is $\tilde{L}_\rho(t_{(r)})$. The NPMLE is obtained as the solution of (12)–(13) for the root u corresponding to the largest value of log-likelihood (4).

As restricted NPMLE's Λ_n, in the class $\mathcal{S}_n \cap \{\Lambda : \Lambda(\tau_0) \leq \tilde{L}_\rho(\tau_0) + 1\}$ defined in corollary 3, exist and are consistent, it must be true a.s. for all sufficiently large n, that such Λ_n lie in the relative interior of $\mathcal{S}_n$ and therefore satisfy (12)–(13). In any case, the NPMLE can now be defined, measurably with respect to ρ, as the estimator of Λ obtained from that admissible solution u of $D(u) = 0$ for which the log-likelihood (4) is largest. For the NPMLE defined in this way, the theorems of section 3 apply to prove consistency.

The algorithm described here is a semiparametric elaboration of a finite-dimensional parametric estimator of Vonta (1996a,b). It has been implemented computationally with good results, as both on real and simulated data we have found generally that over a much longer interval than necessary, the function $D(u)$ for which we are finding a root is monotone increasing.

4.2. *Iterative joint estimation of NPMLE for* (ρ,Λ)

The joint maximization of log-likelihood over (ρ, Λ) also yields consistent estimators. The following theorem is proved in Slud & Vonta (2002) using the theorems above, and is illustrated numerically in the next subsection.

Theorem 4

Under the assumptions of theorem 3, there exists a sufficiently small $\delta > 0$, *such that almost surely as* $n \to \infty$, *there exist joint NPMLE's* $(\hat{\Lambda}_n, \hat{\rho}_n)$ *satisfying* $|\hat{\rho}_n - \rho_0| < \delta$ *and* $\limsup_n \hat{\Lambda}_n(\tau_0) \leq \Lambda_0(\tau_0) + 1$. *For all such joint NPMLE sequences, as* $n \to \infty$ *almost surely* $\hat{\rho}_n \to \rho_0$ *and* $\hat{\Lambda}_n$ *is consistent for* Λ_0 *uniformly on* $[0, \tau_0]$.

4.3. *PBC data example*

We illustrate the joint NPMLEs of the previous section with a frequently analysed clinical trial dataset on PBC (Christensen *et al.*, 1985). The trial consisted of 216 subjects, randomized

either to a placebo or azathioprine treatment group. Survival data were recorded, together with covariates which play no role in our analysis. Of the study subjects, nine had survival times essentially equal to 0 and are excluded from our analysis, and 103 were lost to follow-up before death; 98 of the 207 subjects analysed here were in the treatment group, the remaining 109 on placebo. Previous analyses of this dataset showed a non-significant treatment effect based upon a two-sample logrank statistic, but a good fit to the Cox proportional hazards model based on five to seven covariates and a highly significant (p-value < 0.02) treatment effect after adjusting for these covariates. Here we treat the two-sample data using a Clayton–Cuzick (Gamma-frailty) model with unknown parameter c and treatment-effect parameter ρ. We implemented in Splus the algorithm described in section 4.1 to calculate the restricted NPMLE $\hat{\Lambda}$ for fixed values of the unknown parameters ρ, c. On a grid of approximately 500 pairs ρ, c, within $[1, 1.2] \times [0, 1]$, profiled log-likelihood values were calculated by substituting into formula (4) the restricted NPMLE's $\hat{\Lambda}$ (found to be unique in every case). Using the Splus-supplied bivariate-interpolation and contour functions, we produced the contour plot given as Fig. 1.

As the limiting case of the Clayton–Cuzick model, with parameter c going to 0, is the standard Cox model, previous findings of adequacy of Cox-model fits to the PBC data, agree with the figure showing the profiled PBC log-likelihood to be largest in the neighbourhood of $c = 0$. Thus the two-sample PBC data continue to indicate that Lehmann, or proportional hazards, alternatives fit the data as well as any Clayton–Cuzick frailty model. The parameter ρ maximizing log-likelihood appears to be located near 1.15, although the log-likelihood contours are not nearly sharp enough in that neighbourhood to indicate a significant treatment

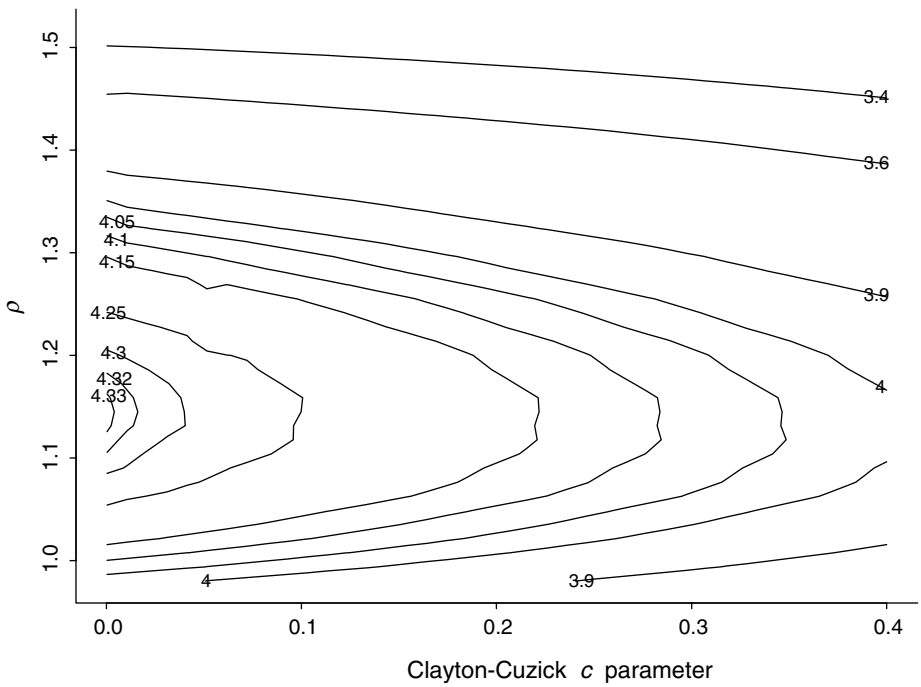

Fig. 1. Smoothed (Splus) contour plot of profiled log-likelihood surface for the PBC data, with respect to parameters ρ and c. The contoured log-likelihood is equal to 590 plus (4) with the maximizer $\hat{\Lambda}$, for fixed (ρ, c), substituted for Λ.

effect. (By analogy with finite-dimensional likelihood theory, ρ significantly different from 1 would be indicated only if the log-likelihood contours were at least $1.92 = \frac{1}{2}(1.96)^2$ lower at $\rho = 1$ than at the maximum near $c = 0$, $\rho = 1.15$.)

5. Extensions

5.1. Nuisance parameters in G

A very interesting extension which can be treated by the methods of this paper is the case where additional unknown finite-dimensional nuisance parameters θ enter the model through $G(\cdot) \equiv G(\cdot, \theta)$. For example, in the model with Gamma-distributed frailty, the scalar parameter $c = \theta$ is generally unknown and must be estimated from the data. Our results carry through in this situation. An analogue of theorem 2 continues to hold, but the preceding algorithm must be modified in order to provide joint NPMLEs of (θ, ρ, Λ). As a first attempt at such an algorithm, we profile the log-likelihood as a function of the finite-dimensional parameters (θ, ρ) by substituting into (4) the maximizer over Λ for fixed (θ, ρ), restricted as in corollary 3, and then optimize in (θ, ρ) by applying a general-purpose function-maximizer on the resulting (spline-smoothed) surface.

5.2. Extension to regression models

The case where the structural parameter is a vector of regression coefficients for observed covariates is of great importance for applications and can be easily handled by our methods. The model is (1) with $\beta \in \mathbf{R}^p$ and z denoting a p-dimensional vector of non-constant explanatory covariates. The log-likelihood (4) evaluated at Λ_ϑ has the same form as before, namely,

$$\sum_z \Big\{ \int \log\Big(e^{\beta' z} G'(e^{\beta' z}\Lambda(t)) \frac{d\Lambda}{dv} \Big) dN_z(t) + \int G(e^{\beta' z}\Lambda(t)) dY_z(t) \Big\},$$

where the sum in z ranges over the distinct observed covariate values. This restriction to finite, as opposed to bounded, support is primarily to enable the use of the explicit preliminary estimator (38) in place of the more complicated density-based estimators of Cheng (1989). Results analogous to the NPML equations (12) and (13) and to theorems 2 and 3 do hold in the regression case. For details, see Slud & Vonta (2002). As an indication of the overall consistency result in this case, we state only one result. In that result, a specific preliminary estimator is cited in equation (38), but any of the others of Cheng *et al*. (1995) would do just as well.

Theorem 5

Denote by β_0 and Λ_0 the true values of the parameters β and Λ governing the data $\{N_z(t), Y_z(t), z \in \mathcal{Z}, t \geq 0\}$ under the model (1) together with (3) and (G.1)–(G.4); and assume that as $n \to \infty$, the group sizes $n_z \equiv Y_z(0)$ grow with n in such a way that a.s. $n_z/n \to c_z > 0$. Assume also that $E(\binom{1}{Z_1}\binom{1}{Z_1}^{\otimes 2})$ is a positive-definite $(p+1) \times (p+1)$ matrix. Define preliminary estimators $(\tilde\beta, \tilde\Lambda)$ through the $(p+1)$-vector equations

$$\sum_z \binom{1}{z} \frac{n_z}{n} S_{KM}^{(z)}(t) = \sum_z \binom{1}{z} \frac{n_z}{n} \exp(-G(e^{z'\tilde\beta}\tilde\Lambda(t))), \tag{38}$$

first by solving (38) at a fixed value t, such as $t = (n^{-1}\sum_z n_z S_{KM}^{(z)})^{-1}(1/2)$, and then by solving the first component equation of (38) for all t. Then any restricted NPMLE of Λ_0 within $S_n \cap \{\Lambda : \Lambda(\tau_0) \leq \tilde\Lambda(\tau_0) + 1\}$, with β fixed at $\tilde\beta$, is a consistent estimator of Λ_0.

6. Discussion

In the general right-censored semiparametric transformation model, we have studied likelihood equations and asymptotic behaviour of the log-likelihood functional in the neighbourhood of the true parameter (β_0, Λ_0) governing a large data-sample. The restricted NPMLEs of Λ for fixed ρ gave information not previously available concerning the profile log-likelihood surface on a non-shrinking neighbourhood of ρ values. This approach differs from the usual one, expounded in Bickel *et al*. (1993), of studying the semiparametric likelihood only through its differential properties at the true parameter point.

As the transformation-model likelihood considered here coincides with that of Nielsen *et al*. (1992), the NPMLEs we studied coincide with theirs. Thus a very useful outcome of our likelihood equations in theorem 1, is the fast algorithm of section 4 for NPML estimation of Λ, in the same setting where Klein (1992) and others apply the EM algorithm. Estimation of ρ and any other unknown parameters such as the constant c in the Clayton–Cuzick model then proceed via a profiled likelihood.

An important direction for further work on the NPMLEs studied in this paper is to establish asymptotic normality and (semiparametric) efficiency (Gill and van der Vaart 1993). Parner (1998) has done this for the specific case of the Clayton–Cuzick model.

The standard idea followed here in proving consistency of NMPLEs has been to characterize expected log-likelihood maximizers uniquely over a sufficiently large subset of parameters within the infinite-dimensionsal parameter space and then to show that, to top order, the log-likelihood and expected log-likelihood have the same maximizers. Two key steps were to establish asymptotic absolute continuity (proposition 1) for restricted NPML sequences Λ_n and to apply the theory of adjoint parameterized ODEs to characterize variationally the maximizer of expected log-likelihood (proposition 2). These steps may prove useful independently in other contexts.

Acknowledgements

This research was partially supported by travel grants from the University of Cyprus and the University of Maryland. We are grateful to the Editors and referees for substantial help in improving the readability of the paper.

References

Bagdonavicius, V. & Nikulin, M. (1997). Analysis of general semiparametric models with random covariates. *Rev. Roumaine Math. Pures Appl.* **42**, 351–369.

Bickel, P., Klaassen, C., Ritov, Y. & Wellner, J. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore.

Cheng, P. (1989). Nonparametric estimator of survival curve under dependent censoring. *J. Statist. Plann. Inference* **23**, 181–192.

Cheng, S. C., Wei, L. J. & Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.

Christensen, E., Neuberger, J., Crowe, J., Altman, D., Popper, H., Portmann, B., Doniach, D., Ranek, L., Tygstrup, N. & Williams, R. (1985). Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis. *Gastroenterology* **89**, 1084–1091.

Clayton, D. & Cuzick, J. (1986). The semi-parametric Pareto model for regression analysis of survival times. *Papers on semiparametric models at the ISI centenary session, Amsterdam*, Report MS-R8614, Centrum voor Wiskunde en Informatica, Amsterdam.

Coddington, E. & Levinson, N. (1957). *Theory of ordinary differential equations*. McGraw-Hill, New York.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187–202.

Dabrowska, D. & Doksum, K. (1988). Partial likelihood in transformation models with censored data. *Scand. J. Statist.* **15**, 1–23.
Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1). *Scand. J. Statist.* **16**, 97–128.
Gill, R. (1992). Marginal partial likelihood. *Scand. J. Statist.* **79**, 133–137.
Gill, R. D. & van der Vaart, A. (1993). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 2). *Scand. J. Statist.* **20**, 271–288.
Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika* **73**, 387–396.
Johansen, S. (1983). An extension of Cox's regression model. *Internat. Statist. Rev.* **51**, 165–174.
Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.* **27**, 887–906.
Klaassen, C. (1993). *Efficient estimation in the Clayton–Cuzick model for survival data*. Preprint, University of Amsterdam.
Klein, J. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795–806.
Murphy, S. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.* **22**, 712–731.
Murphy, S. (1995). Asymptotic theory for the frailty model. *Ann. Statist.* **23**, 182–198.
Murphy, S. & van der Vaart, A. (2000). On profile likelihood, *J. Amer. Statist. Assoc.* **95**, 449–465.
Nielsen, G., Gill, R., Andersen, P. & Sørensen, T. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.* **19**, 25–44.
Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.* **26**, 183–214.
Slud, E. (1992). Partial likelihood for continuous-time stochastic processes. *Scand. J. Statist.* **19**, 97–110.
Slud, E. & Vonta, F. (2002). *Nonparametric likelihood and consistency of NPMLE's in the transformation model*. Technical Report TR/17/02, Mathematics and Statistics Department, University of Cyprus.
Vonta, F. (1992). *Efficient estimation of a structural parameter in a non-proportional hazards model in the two-sample problem*. Mathematics Department Thesis, University of Maryland.
Vonta, F. (1996a). Estimation in two-sample nonproportional hazards models in clinical trials by an algorithmic method. In *Proceedings in Computational Statistics* (ed. A. Prat). Physica-Verlag Heidelberg, pp. 489–495.
Vonta, F. (1996b). *Efficient estimation in a nonproportional hazards model in survival analysis. Scand. J. Statist.* **23**, 49–62.

Received February 2002, in final form March 2003

Eric V. Slud, Mathematics Department, University of Maryland, College Park, MD 20742, USA.
E-mail: evs@math.umd.edu

Appendix A: Derivative of logLik

Lemma 3

Let $\hat{\Lambda}$ be an NPMLE in the space $\mathcal{S}_0$, for fixed ρ, and let $J = \{t : \Delta\hat{\Lambda}(t) > 0\}$ be the countable set of its jump points. Then $\hat{\Lambda}$ is a pure jump function; all jumps of N occur at jumps of $\hat{\Lambda}$; and for $t \in J$,

$$\frac{\Delta N(t)}{\Delta\hat{\Lambda}(t)} + \sum_z \rho^z \Big\{ (-Y_z(t)) G'(\rho^z \hat{\Lambda}(t)) + \frac{G''(\rho^z \hat{\Lambda}(t))}{G'(\rho^z \hat{\Lambda}(t))} \Delta N_z(t) + \int_t^\infty \Big((-Y_z(s)) d(G' \circ \rho^z \hat{\Lambda})(s) + \frac{G''(\rho^z \hat{\Lambda}(s))}{G'(\rho^z \hat{\Lambda}(s))} \Delta N_z(s) \Big) \Big\} = 0. \tag{39}$$

Proof. The steps of the proof are only summarized here, with full details in Slud & Vonta (2002). We successively set to 0 the first Gâteaux derivative of the log-likelihood function, given in (9), for specific choices of γ. First, with $\gamma = I_{[\Delta N(t) \neq 0, \Delta\hat{\Lambda}(t) = 0]}$, the derivative equation (9) implies that $I_{[\Delta\hat{\Lambda}(t)=0]} dN(t)$ is the 0-measure. Next, with $\gamma = I_{[\Delta\hat{\Lambda}(t)=0]}$, equation (9) reduces

after integration by parts and some algebra, to show via (G.1) that $I_{[\Delta\hat{\Lambda}(s)=0]}d\hat{\Lambda}(s)$ is also the 0-measure.

The remaining non-zero terms in the differentiated likelihood (9) lead, after changing the order of integration in the terms involving $\int_0^{t-}\gamma d\hat{\Lambda}$ and noting $I_{[\Delta\hat{\Lambda}(t)>0]} = 1$ a.e. $(d\hat{\Lambda}(t))$, to

$$\sum_z \left\{ \int \gamma(t) \left(\frac{\Delta N_z(t)}{\Delta\hat{\Lambda}(t)} + \rho^z \left\{ \left((-Y_z(t)) G'(\rho^z\hat{\Lambda}(t)) + \frac{G''(\rho^z\hat{\Lambda}(t))}{G'(\rho^z\hat{\Lambda}(t))} \Delta N_z(t) \right) \right.\right.$$
$$\left.\left. + \int_t^\infty \left((-Y_z(s)) d(G'\circ\rho^z\hat{\Lambda})(s) + \frac{G''(\rho^z\hat{\Lambda}(s))}{G'(\rho^z\hat{\Lambda}(s))} \Delta N_z(s) \right) \right\} \right) d\hat{\Lambda} \right\} = 0.$$

The last equation must hold for all bounded measurable functions γ, implying (39) and completing the proof.

Lemma 4
For an NPMLE $\hat{\Lambda} \in \mathcal{S}_0$, $\{t : \Delta\hat{\Lambda}(t) > 0\} = \{t : \Delta N(t) > 0\}$.

Proof. By lemma 3, $\hat{\Lambda}$ is a pure-jump function. Integration by parts in (39) implies

$$\frac{\Delta N(t)}{\Delta\hat{\Lambda}(t)} + \sum_z \rho^z \left\{ \frac{G''(\rho^z\hat{\Lambda}(t))}{G'(\rho^z\hat{\Lambda}(t))} \Delta N_z(t) \; + \int_t^\infty \left((G'\circ\rho^z\hat{\Lambda})(s) dY_z(s) + \frac{G''(\rho^z\hat{\Lambda}(s))}{G'(\rho^z\hat{\Lambda}(s))} \Delta N_z(s) \right) \right\} = 0.$$

If $\hat{\Lambda}$ had a jump at $t < \tau_0$ and $\Delta N(t) = 0$, then only the integral in the last equation remains. The integrand is ≤ 0 by (G.1), while by (3), $\sum_z G'(\rho^z\hat{\Lambda}(\tau_0))\Delta Y_z(\tau_0) < 0$. This contradiction completes the proof.

Proof of theorem 1. Let $s < t$ be two successive jump points of $\hat{\Lambda}$. By subtracting (39) taken at s from (39) taken at t, we get

$$\frac{\Delta N(t)}{\Delta\hat{\Lambda}(t)} + \sum_z \rho^z \left\{ -Y_z(t) G'(\rho^z\hat{\Lambda}(t)) + \frac{G''(\rho^z\hat{\Lambda}(t))}{G'(\rho^z\hat{\Lambda}(t))} \Delta N_z(t) \right.$$
$$- \frac{\Delta N(s)}{\Delta\hat{\Lambda}(s)} - Y_z(s) G'(\rho^z\hat{\Lambda}(s)) - \frac{G''(\rho^z\hat{\Lambda}(s))}{G'(\rho^z\hat{\Lambda}(s))} \Delta N_z(s)$$
$$\left. - \int_{s^+}^t \left(-Y_z(x) d(G'\circ\rho^z\hat{\Lambda})(x) + \frac{G''(\rho^z\hat{\Lambda}(x))}{G'(\rho^z\hat{\Lambda}(x))} \Delta N_z(x) \right) \right\} = 0.$$

As s and t are successive jumps of $\hat{\Lambda}$, the last term above is equal to

$$\sum_z \rho^z \left\{ -Y_z(t) \Delta(G'\circ\rho^z\hat{\Lambda})(t) + \frac{G''(\rho^z\hat{\Lambda}(t))}{G'(\rho^z\hat{\Lambda}(t))} \Delta N_z(t) \right\}.$$

After some obvious cancellations we obtain equation (12).

Now let t_* be the last jump point of $\hat{\Lambda}$. Then for $t = t_*$, equation (39) leads to equation (13), as there are no jumps of $\hat{\Lambda}$ after t_*, making the integral from t_* to ∞ equal to 0.

Appendix B: Proofs of proposition 1, theorem 2

Proof of proposition 1. The idea of the proof is to define, for sequences of $\Lambda_n \in \mathcal{S}_n$ satisfying $\Lambda_n(\tau_0) \le K$ such that (with positive probability, for arbitrarily large n and) for some $j \le m$,

$$\frac{\Lambda_n(\gamma_{j+1}) - \Lambda_n(\gamma_j)}{\Lambda_0(\gamma_{j+1}) - \Lambda_0(\gamma_j)} > C, \tag{40}$$

a new sequence $\tilde{\Lambda} \equiv \tilde{\Lambda}_n \le \Lambda_n$ in $\mathcal{S}_n$ such that (for all large n)

$$\text{logLik}(\tilde{\Lambda}, \rho) > \text{logLik}(\Lambda_n, \rho),$$

where the log-likelihoods throughout this proposition and throughout the paper are all calculated for the fixed value ρ, not for ρ_0. The sequence $\tilde{\Lambda}$ is defined at all jump-points t of $N(\cdot)$, with j fixed satisfying (40), by

$$\Delta\tilde{\Lambda}(t) = \begin{cases} \Delta\Lambda_n(t) & \text{for } t \notin (\gamma_j, \gamma_{j+1}) \\ b\Delta\Lambda_n(t) & \text{for } t \in (\gamma_j, \gamma_{j+1}) \end{cases},$$

for an arbitrary positive constant $b \in (1/2,1)$. Note that $\tilde{\Lambda}$ so defined does satisfy $\tilde{\Lambda}(\tau_0) \le \Lambda_n(\tau_0) \le K$.

Using this definition, (14) and (40) with

$$C = 2G'(0)(1-b)^{-1} \log\left(\frac{1}{b}\right) \frac{\sum_z \rho_0^z}{\sum_z \rho^z G'(\rho^z K) q_z(\tau_0)},$$

straightforward estimates – full details of which can be found in Slud & Vonta (2002) – show that for all large n, and arbitrarily small $1 - b$,

$$\text{logLik}(\tilde{\Lambda}, \rho) - \text{logLik}(\Lambda_n, \rho) > 0.$$

Thus Λ_n cannot have been the NPMLE, and the proof is complete.

Proof of theorem 2. By lemma 2 for $\Lambda = \Lambda_n \in \mathcal{S}_n$,

$$\frac{1}{n}\text{logLik}(\Lambda, \rho) \le \frac{1}{n}\sum_z \int (G \circ \rho^z \Lambda) dY_z + \sum_z \sum_j \frac{r_{j,z}}{n} \log \frac{C_{j,z}(\Lambda)}{r_{j,z}}.$$

The Glivenko–Cantelli lemma implies that $Y_z(\cdot)/n$ converges a.s., uniformly as $n \to \infty$, to non-increasing continuous q_z. Therefore,

$$\lim_n \sum_z \int (G \circ \rho^z \Lambda) \frac{dY_z}{n} = \sum_z \int (G \circ \rho^z \Lambda) dq_z.$$

The remainder of the proof, which uses proposition 1, the asymptotic smallness of $\max_{i \le m}(\Lambda_n(\gamma_{i+1}) - \Lambda_n(\gamma_i))$ and the information inequality (16), can be found in Slud & Vonta (2002).

Appendix C: Justification of (31)

The adjoint system (29)–(30) is a linear matrix equation

$$\begin{pmatrix} dL_*/d\Lambda_0 \\ dP_*/d\Lambda_0 \end{pmatrix} = A \begin{pmatrix} L_* \\ P_* \end{pmatrix}, \quad \begin{pmatrix} L_*(0) \\ P_*(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{41}$$

with the entries $A_{ij}(s)$ of the 2×2 matrix $A(s)$ given by

$$A_{11}(s) = -A_{22}(s) = -\frac{\sum_z \rho_0^{2z} q_z(s) G''(\rho_0^z \Lambda_0(s))}{\sum_z \rho_0^z q_z(s) G'(\rho_0^z \Lambda_0(s))}, \quad A_{12}(s) = -\frac{1}{\sum_z \rho_0^z q_z(s) G'(\rho_0^z \Lambda_0(s))},$$

and

$$\begin{aligned} A_{21}(s) &= \frac{\left(\sum_z \rho_0^{2z} q_z(s) G''(\rho_0^z \Lambda_0(s))\right)^2}{\sum_z \rho_0^z q_z(s) G'(\rho_0^z \Lambda_0(s))} - \sum_z \rho_0^{3z} q_z(s) \left(\frac{G''^2}{G'}\right)_{\rho_0^z \Lambda_0(s)} \\ &= -\rho_0 q_1(s) q_0(s) \frac{(G'(\rho_0 \Lambda_0(s)) G''(\Lambda_0(s)) - \rho_0 G''(\rho_0 \Lambda_0(s)) G'(\Lambda_0(s)))^2}{G'(\rho_0 \Lambda_0(s)) G'(\Lambda_0(s)) \sum_z \rho_0^z q_z G'(\rho_0^z \Lambda_0)}. \end{aligned}$$

By inspection of the foregoing definitions, $A_{21}(s) < 0$, $A_{11}(s) > 0$, $A_{12}(s) < 0$ uniformly on each interval $[\epsilon, \tau_0]$ for which $\Lambda_0(\tau_0) > 0$, and

$$\mathrm{tr}(A(s)) = 0, \quad \det(A(s)) = -\frac{\sum_z \rho_0^{3z} q_z(s) (G''^2/G')_{\rho_0^z \Lambda_0(s)}}{\sum_z \rho_0^z q_z(s) G'(\rho_0^z \Lambda_0(s))}.$$

Then if we define

$$B(s) = \int_0^s A(u) d\Lambda_0(u) = \begin{pmatrix} \xi & -\eta \\ (\xi^2 - \Delta^2)/\eta & -\xi \end{pmatrix},$$

we have

$$\det(B) = -\Delta^2, \quad \xi, \Delta, \eta > 0, \ \ \xi < \Delta,$$

with all inequalities uniform over $[\epsilon, \tau_0]$, where

$$\xi(s) = \int_0^s A_{11}(u) d\Lambda_0(u), \quad \eta(s) = -\int_0^s A_{12}(u) d\Lambda_0(u),$$

and

$$\Delta^2(s) = \xi^2(s) + \eta(s) \int_0^s \frac{\det A(u) + A_{11}^2(u)}{A_{12}(u)} d\Lambda_0(u).$$

It is easily checked that $(\eta, \xi + \Delta)'$ and $(\eta, \xi - \Delta)'$ are, respectively, right eigenvectors for B with eigenvalues $-\Delta$ and Δ, so that

$$B = \begin{pmatrix} \eta & \eta \\ \xi + \Delta & \xi - \Delta \end{pmatrix} \begin{pmatrix} -\Delta & 0 \\ 0 & \Delta \end{pmatrix} \begin{pmatrix} \eta & \eta \\ \xi + \Delta & \xi - \Delta \end{pmatrix}^{-1},$$

and

$$\begin{aligned} \exp(B) &= \frac{1}{2\eta\Delta} \begin{pmatrix} \eta & \eta \\ \xi + \Delta & \xi - \Delta \end{pmatrix} \begin{pmatrix} e^{-\Delta} & 0 \\ 0 & e^{\Delta} \end{pmatrix} \begin{pmatrix} \Delta - \xi & \eta \\ \Delta + \xi & -\eta \end{pmatrix} \\ &= \begin{pmatrix} \cosh(\Delta) + (\xi/\Delta)\sinh(\Delta) & -(\eta/\Delta)\sinh(\Delta) \\ ((\xi^2 - \Delta^2)/\eta\Delta)\sinh(\Delta) & \cosh(\Delta) - (\xi/\Delta)\sinh(\Delta) \end{pmatrix}. \end{aligned}$$

It follows immediately that $e^B = \cosh(\Delta)\mathbf{I} + (\sinh(\Delta)/\Delta)B$, and if $u_1 \le 0$, $u_2 > 0$, then

$$e^B \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \in \left\{ \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} : v_1 \le 0, v_2 \ge e^{-\Delta} u_2 \right\}. \tag{42}$$

Now the linear matrix ODE (41) has a unique solution, which can be expressed as the limit over $h \to 0$ of the solutions of the approximating linear system in which the coefficients $A(s)$ are

replaced by the piecewise-constant function equal to $A(kh)$ on the interval $s \in [kh, (k+1)h)$. That is, in terms of the matrices $B(s)$ defined above, the solution on $[0, \tau_0]$ is

$$\begin{pmatrix} L_*(s) \\ P_*(s) \end{pmatrix} = \lim_{h \searrow 0} \prod_{k \geq 0: kh \leq s} e^{B((k+1)h) - B(kh)} \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

where the multiplication in the product is done in the order with lowest-index terms furthest to the right. Then by (42), we conclude that $P_*(s) \geq \exp(-\int_0^s (-\det(A(x))^{1/2} d\Lambda_0(x))$. This integral is bounded away from 0 on $[0, \tau_0]$ because the formula given above for $\det(A(x))$, shows that $(-\det(A(x))^{1/2}$ is uniformly bounded above on $[0, \tau_0]$.

Thus $P(\tau_0, \alpha, \rho)$ is a well-defined and continuously differentiable function of (α, ρ) on a neighbourhood of the point $(0, \rho_0)$ at which $P(\tau_0, 0, \rho_0)$ is 0, and (31) follows.

Appendix D: Proof of theorem 3

Proposition 1 has already shown that for fixed ρ and $K > \Lambda_0(\tau_0)$, an NPMLE sequence exists satisfying $\limsup_n \Lambda_n(\tau_0) \leq K$. The main assertion of the theorem will follow when we exhibit an element $\Lambda_{\rho,n} \in \mathcal{S}_n$, which is not necessarily an NPMLE but which satisfies (34) and $\limsup_n \Lambda_{\rho,n}(\tau_0) \leq K < \infty$. In terms of the partitioning sequence $\{\gamma_{in}\}$ defined in (P.1), and the maximizing function L_ρ found in proposition 2, define the random element $\Lambda_{\rho,n} \in \mathcal{S}_n$ by:

$$\Delta\Lambda_{\rho,n}(t) \equiv \frac{1}{r_{j*}} (L_\rho(\gamma_{j+1}) - L_\rho(\gamma_j)) \quad \text{for} \quad t \in (\gamma_j, \gamma_{j+1}], \ \Delta N(t) > 0. \tag{43}$$

By continuity of Λ_0, there are a.s. no times t which are simultaneously jumps of N_0 and N_1. Clearly the functions $\Lambda_{\rho,n}$ and L_ρ agree at all points $\gamma_j, j = 0, \ldots, m+1$, and for fixed $K > \Lambda_0(\tau_0)$ and ρ in a sufficiently small neighbourhood of ρ_0, the proof of proposition 2 shows that $L_\rho(\tau_0) < K$, which now implies that $\Lambda_{\rho,n}(\tau_0) < K$. Recalling that the discrete part of v is the counting measure of jumps of N, we find from the definitions (43) and (4) that

$$\begin{aligned} \frac{1}{n}\left[\log\mathrm{Lik}(\Lambda_{\rho,n}, \rho) + N(\infty)\log n\right] &= \frac{1}{n}\sum_z \int G \circ \rho^z \Lambda_{\rho,n} dY_z \\ &+ \frac{1}{n}\sum_{z,j} \int_{\gamma_j}^{\gamma_{j+1}} \log\left(\frac{n}{r_{j*}} \rho^z G'(\rho^z \Lambda_{\rho,n}(t))(L_\rho(\gamma_{j+1}) - L_\rho(\gamma_j))\right) dN_z(t), \end{aligned}$$

which by (17) and (18) a.s. has limiting value $\mathcal{J}(L_\rho, \rho)$ as $n \to \infty$.

As $\Lambda_{\rho,n}$ is a random element of $\mathcal{S}_n$ with $\limsup_n \Lambda_{\rho,n}(\tau_0) \leq K$, with $\Lambda_{\rho,n} \longrightarrow L_\rho$ uniformly on $[0, \tau_0]$, it follows that for sequences Λ_n of maximizers of (4) within $\{\Lambda \in \mathcal{S}_n : \Lambda(\tau_0) \leq K\}$, $n^{-1}(\log\mathrm{Lik}(\Lambda_n) + N(\infty)\log n)$ must asymptotically on the one hand be at least as large as $\mathcal{J}(L_\rho, \rho)$, and on the other hand (by theorem 2) can be no larger than $\mathcal{J}(L_\rho, \rho)$. Thus (34) holds.

If (35) did not also hold, then with positive probability, for some $\epsilon > 0$, a subsequence $\Lambda_{n'}$ would fall in $A = \{\Lambda : \sup_{t \in [0, \tau_0]} |\Lambda(t) - L_\rho(t)| \geq \epsilon\}$. By corollary 1, the quantities $(n')^{-1} \cdot (\log\mathrm{Lik}(\Lambda_{n'}) + N(\infty)\log n')$ could be at most $\sup_{L \in A} \mathcal{J}(L, \rho)$, which by proposition 2 is strictly smaller than $\mathcal{J}(L_\rho, \rho)$. This contradiction proves (35) and the theorem.