Statistica Sinica **18**(2008), 467-492

ESTIMATING MONOTONE, UNIMODAL AND U–SHAPED FAILURE RATES USING ASYMPTOTIC PIVOTS

Moulinath Banerjee

University of Michigan

Abstract: We propose a new method for pointwise estimation of monotone, unimodal and U–shaped failure rates, under a right–censoring mechanism, using nonparametric likelihood ratios. The asymptotic distribution of the likelihood ratio is pivotal, though non–standard, and can therefore be used to construct asymptotic confidence intervals for the failure rate at a point of interest, via inversion. Major advantages of the new method lie in the facts that it completely avoids estimation of nuisance parameters, or the choice of a bandwidth/tuning parameter, and is extremely easy to implement. The new method is shown to perform competitively in simulations, and is illustrated on a data set involving time to diagnosis of schizophrenia in the Jerusalem Perinatal Cohort Schizophrenia Study.

Key words and phrases: Asymptotic pivots, greatest convex minorant, likelihood ratio statistic, monotone hazard rate, two–sided Brownian motion, universal limit.

1. Introduction

The study of hazard functions arises naturally in lifetime data analysis, a key topic of interest in reliability and biomedical studies. By "lifetime" we usually mean the time to failure/death, infection, or the development of a syndrome of interest, usually assumed random. While a random variable is typically characterized by its density function or distribution function, in the study of lifetimes the instantaneous hazard function/failure rate is often a more useful way of describing the behavior of the random variable. If F denotes the distribution function of the lifetime of an individual, then the cumulative hazard function, given by $\Lambda = -\log(1-F)$, is increasing and assumes values in $[0, \infty)$. The instantaneous hazard rate, denoted by λ is the derivative of Λ; thus, $\lambda(x) = f(x)/(1-F(x))$. This quantity is called the instantaneous hazard function/failure rate; a higher value of $\lambda(x)$ indicates a greater chance of failure in the instant after x. Information on $\lambda(x)$ is of vital importance to reliability engineers/ medical practitioners, since, among other things, it enables them to gauge the necessity of adopting some mode of preventive action/intervention at any particular time to keep the system/patient from failing.

In many applications, one can impose natural qualitative constraints on the failure rate, in terms of shape restrictions. The most important kinds of shape restrictions are monotonicity (increasing or decreasing) and bathtub/U shapes. Human life, for example, can be appropriately described by a bathtub-shaped hazard. The initial period is high–risk (owing to the risk of birth defects and infant disease), followed by a period of more or less constant risk, and then the risk starts increasing again, due to the onset of the aging process. Models with increasing or decreasing hazards are also fairly common. An example, given by Gijbels and Heckman (2004), comes from an industry where manufacturers use a "burn–in" process for their products. The products are subjected to operation before being sold to customers. This helps in preventing an early failure of defective items, and the robust ones that are put on the market subsequently exhibit gradual aging (increasing failure rate with time). Decreasing hazards are useful for modelling survival times after a successful medical treatment. If the operation/therapy is useful and of long term consequence, then the risk of failure from the condition that required the therapy should go down over an appreciably long time interval following treatment.

Though shape constrained hazards appear quite frequently in many important applications, as noted above, there are relatively few nonparametric methods for estimating hazard rates and, in particular, for constructing reliable confidence intervals for the hazard rate under shape constraints. Confidence sets for the hazard function are important: they are more informative than the point estimate in that they provide a range of plausible values for the hazard rate at the point of interest, and can be used more effectively for decision making. In this paper, we propose a novel method, using asymptotic pivots, for constructing nonparametric pointwise confidence sets for a monotone failure rate. We then extend this idea to the study of unimodal or U–shaped hazards. While our discussion is framed in the realistic context of right–censored data, the proposed methodology is equally applicable to uncensored data.

2. Background

In the uncensored case we observe $X_1, \ldots, X_n$, i.i.d. lifetimes with distribution function F (concentrated on $(0, \infty)$) and density function f, and the goal is to estimate $\lambda(x) = f(x)/(1 - F(x))$ based on these data. In the case of right–censored data, not all the X_i's are observable; rather, one observes pairs of random variables $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ where $\delta_i = 1\{X_i \leq Y_i\}$ and $T_i = X_i \wedge Y_i$ where Y_i is the (random) time of observation of the i'th individual. It is assumed that the Y_i's are mutually independent and independent of the X_i's. For an individual, $\delta_i = 1$ implies that they have been observed to fail, and that the exact time to failure is known. On the other hand $\delta_i = 0$ implies that failure was not

observed during the observation time period and the information on failure time is therefore censored. The goal is to estimate λ.

Maximum likelihood estimators for an increasing hazard function based on uncensored i.i.d. data were studied by Grenander (1956) and Marshall and Proschan (1965), and the asymptotic distribution of the MLE at a fixed point in the uncensored case was studied by Prakasa Rao (1970). Padgett and Wei (1980) derived the MLE of an increasing hazard function based on right–censored data but without further development of its asymptotic properties. Mykytyn and Santner (1981) also studied non-parametric maximum likelihood estimation based on monotonicity assumptions concerning the hazard rate λ, and several different censoring schemes. See also Wang (1986), Tsai (1988) and Mukerjee and Wang (1993) for maximum likelihood based estimation for an increasing hazard function. With right–censored data, the asymptotic distribution of the MLE of a monotone increasing hazard λ at a fixed point was derived by Huang and Wellner (1995). This result is connected with the development in this paper, and we return to it in more detail later. Huang and Wellner (1995) showed that the rate of convergence of $\hat{\lambda}(t)$, the MLE of λ evaluated at the point t is $n^{1/3}$; more precisely, at a fixed point t_0, $n^{1/3}\,(\hat{\lambda}(t_0) - \lambda(t_0)) \to C(t_0)\,2\,\mathbb{Z}$, under modest assumptions. In particular, the assumption of strict monotonicity of λ at the point t_0 ($\lambda'(t_0) > 0$) is needed for the $n^{1/3}$ rate of convergence to hold. Here, $C(t_0)$ is a constant depending on t_0 and the underlying parameters of the problem, and $\mathbb{Z} = \text{argmin}_{h\in\mathbb{R}}\,(W(h) + h^2)$ with $W(h)$ being two–sided Brownian motion starting from 0.

The above result yields a method for constructing a confidence set for $\lambda(t_0)$. If $\hat{C}(t_0)$ is a consistent estimator for $C(t_0)$, a large sample level $1-\alpha$ confidence interval for $\lambda(t_0)$ is given by $[\hat{\lambda}(t_0) - n^{-1/3}2\hat{C}(t_0)\,q(\mathbb{Z}, 1-\alpha/2), \hat{\lambda}(t_0) + n^{-1/3}2\hat{C}(t_0)\,q(\mathbb{Z}, 1-\alpha/2)]$, where $q(\mathbb{Z}, 1-\alpha/2)$ is the $(1-\alpha/2)$'th quantile of the distribution of $\mathbb{Z}$. Quantiles of the distribution of $\mathbb{Z}$ are tabulated in Groeneboom and Wellner (2001). The main difficulty with this confidence set is that of estimating the nuisance parameter $C(t_0)$ which, among other things, depends on $\lambda'(t_0)$. Estimating the derivative of the hazard in this setting is a tricky affair. One option is to kernel smooth the MLE $\hat{\lambda}$; this turns out to be more complex in comparison to the the kernel smoothing procedures employed in density estimation based on i.i.d. observations from an underlying distribution, or in nonparametric regression. In contrast to the standard density estimation case, the number of support points of $\hat{\lambda}$ is of a smaller order; consequently, direct kernel smoothing with naive bandwidth choices may not recover all the information lost in the discrete NPMLE of λ. Selection of an optimal bandwidth in terms of the bias–variance tradeoff that is standard for the usual density estimation/nonparametric regression scenarios is also not a realistic option here, since

convenient expressions for the bias and the variance are much harder to compute in this case (and in similar models, involving nonparametric maximum likelihood estimation of a monotone function). Recent work by Groeneboom and Jongbloed (2003) in a related model suggests that a bandwidth of order $n^{-1/7}$ may be appropriate in this context while other options include bandwidth selection based on cross–validation techniques, or the use of the derivative of a standard kernel based estimator of the instantaneous hazard function, ignoring the monotonicity constraint However, the last option is somewhat ad–hoc, and not completely desirable because of its failure to guarantee a slope of the right sign. In summary, estimation of $\lambda'(t_0)$ is not an easy problem and can be heavily influenced by the choice of the bandwidth, thereby introducing variability into the constructed confidence interval.

Recently, Hall, Huang, Gifford and Gijbels (2001) have proposed smooth estimates of the instantaneous hazard under the constraint of monotonicity, with uncensored and both right censored data. With uncensored data, they construct a kernel type estimate of the underlying density function, with different probabilities assigned to the different observations, and choose that probability vector that minimizes a distance measure from the vector of uniform probabilities, subject to maintaining that the hazard function corresponding to the kernel estimate of the density is always non–negative/non–positive according to the constraint. This minimizing vector is then used to compute the proposed estimates of the density, distribution function and hazard. The procedure extends naturally to the right–censored case. Their method imposes constraints at an infinite number of points, and to employ it in practice, they discretize to a very fine grid and resort to quadratic programming routines. This provides yet another route to estimating the derivative of the hazard function, but still requires the choice of bandwidth. While the emphasis in Hall et al. (2001) is to propose a new smooth estimate of the monotone instantaneous hazard on its domain, they also indicate how pointwise confidence bands may be constructed at the end of their Section 2, by using the asymptotic normality of the proposed estimate of the hazard. However, they do not provide a detailed discussion of the nature or reliability of the confidence sets thus obtained (not surprisingly, as the focus of the paper is somewhat different) though they note that their proposed bounds do not account for the bias component of the estimator $\hat{\lambda}$. They also note that this problem can be alleviated by substantial undersmoothing when computing $\hat{\lambda}$, in which case the bands will widen substantially (and therefore become less informative), or by directly estimating bias, which is not really practicable.

In this paper, we provide a new method for constructing pointwise confidence sets for a monotone hazard rate that extends readily to unimodal/U–shaped hazards and dispenses with some of the issues of the existing methods. Our method

is based on inversion of the likelihood ratio statistic for testing the value of a monotone hazard at a point. The likelihood ratio statistic is shown to be asymptotically pivotal with a known limit distribution, whence confidence sets may be obtained by regular inversion, with calibration provided by the quantiles of the limit distribution. As we will see later, good numerical approximations to these quantiles are well–tabulated and hence can be readily used. The most attractive features of the proposed method are (i) it does not involve estimating nuisance parameters, or the choice of a smoothing parameter, and in that respect is more automated and objective than competing methods; (ii) it is computationally inexpensive, as it requires only elementary applications of the PAVA (pool adjacent violators algorithm) or a standard isotonic regression algorithm. To our knowledge, this is the first method in the literature on hazard function estimation under shape constraints that completely does away with the estimation of nuisance parameters or tuning parameters. It must be noted, however, that we are not lead to a new estimator of the instantaneous hazard (unlike the method proposed by Hall et al. (2001)), as it is based on unconstrained and constrained MLE's of λ. The use of inversion of the likelihood ratio statistic to construct a confidence set for the hazard function at a point is motivated by recent developments in likelihood ratio inference for monotone functions initiated in the context of current status data by Banerjee and Wellner (2001), and investigated more thoroughly in the context of conditionally parametric models by Banerjee (2007).

The rest of the paper is organized as follows. In Section 3, we describe the likelihood ratio method for estimating a monotone hazard, and show how the methodology extends to the study of unimodal or U–shaped hazards. Section 4 presents results from simulation experiments and illustrates the new methodology on a dataset on time to development of schizophrenia. Section 5 concludes with a brief discussion of some of the open problems in this area. Proofs and proof–sketches of some of the main results are presented in Section 6 (the appendix) which is followed by references.

Before proceeding to the next section, we introduce the stochastic processes and derived functionals that are needed to describe the asymptotic distributions. We first need some notation. For a real–valued function f defined on $\mathbb{R}$, let $\mathrm{slogcm}(f, I)$ denote the left–hand slope of the GCM (greatest convex minorant) of the restriction of f to the interval I. We abbreviate $\mathrm{slogcm}(f, \mathbb{R})$ to $\mathrm{slogcm}(f)$. Take

$$\mathrm{slogcm}^0(f) = (\mathrm{slogcm}\,(f, (-\infty, 0]) \wedge 0)\, 1_{(-\infty,0]} + (\mathrm{slogcm}\,(f, (0, \infty)) \vee 0)\, 1_{(0,\infty)}\,.$$

For positive constants c and d define the process $X_{c,d}(z) = c\,W(z) + d\,z^2$, where $W(z)$ is standard two-sided Brownian motion starting from 0. Set $g_{c,d} =$

slogcm$(X_{c,d})$ and $g^0_{c,d} = \text{slogcm}^0\,(X_{c,d})$. It is known that $g_{c,d}$ is a piecewise constant increasing function, with finitely many jumps in any compact interval. Also $g^0_{c,d}$, like $g_{c,d}$, is a piecewise constant increasing function with finitely many jumps in any compact interval and differing, almost surely, from $g_{c,d}$ on a finite interval containing 0. In fact, with probability 1, $g^0_{c,d}$ is identically 0 in some random neighbourhood of 0, whereas $g_{c,d}$ is almost surely non-zero in some random neighbourhood of 0. The length of the interval $D_{c,d}$ on which $g_{c,d}$ and $g^0_{c,d}$ differ is $O_p(1)$. For more detailed descriptions of the processes $g_{c,d}$ and $g^0_{c,d}$, see Banerjee and Wellner (2001) and Wellner (2003). Thus, $g_{1,1}$ and $g^0_{1,1}$ are the unconstrained and constrained versions of the slope processes associated with the canonical process $X_{1,1}(z)$. By Brownian scaling, the slope processes $g_{c,d}$ and $g^0_{c,d}$ can be related in distribution to the canonical slope processes $g_{1,1}$ and $g^0_{1,1}$. The following lemma holds. For positive a and b, set $\mathbb{D}_{a,b} = \int \{(g_{a,b}(u))^2 - (g^0_{a,b}(u))^2\}\, du$. Abbreviate $\mathbb{D}_{1,1}$ to $\mathbb{D}$.

Lemma 2.1. *For positive a and b, $\mathbb{D}_{a,b}$ has the same distribution as $a^2\,\mathbb{D}$.*

See Banerjee and Wellner (2001).

3. The Likelihood Ratio Method

In what follows we first assume that the hazard function is monotone increasing, and we discuss the more general case of right censored data. We are concerned with the asymptotic distribution of the likelihood ratio statistic (LRS) for testing the null hypothesis $\lambda(t_0) = \theta_0$, where t_0 is some pre–fixed interior point in the domain of f.

The model with right censoring: Here we have n underlying i.i.d. pairs of non–negative random variables, $(X_1, Y_1), \ldots, (X_n, Y_n)$, with X_i independent of Y_i. We can think of X_i as the survival time of the i'th individual and of Y_i as the time observed. We observe $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ where $\delta_i = 1\{X_i \le Y_i\}$ and $T_i = X_i \wedge Y_i$. Denote the distribution of the survival time by F and the distribution of the observation time by K. The distribution of $T = X \wedge Y$ is denoted by H and relates to F and K in the following way:

$$\overline{H}(x) = (1 - F(x))\,(1 - K(x)) \equiv \overline{F}(x)\,\overline{K}(x)\,.$$

With $\mathcal{D} \equiv ((T_1, \delta_1), \ldots, (T_n, \delta_n))$, the likelihood function for the data can be written as:

$$\begin{aligned} L_n(\mathcal{D}, \lambda) &= \Pi_{i=1}^n \left(f(T_i)\,\overline{K}(T_i)\right)^{\delta_i} \left(g(T_i)\,\overline{F}(T_i)\right)^{1-\delta_i} \\ &= \Pi_{i=1}^n \left(\frac{f(T_i)}{\overline{F}(T_i)}\right)^{\delta_i} \overline{F}(T_i) \times \Pi_{i=1}^n\, \overline{K}(T_i)^{\delta_i}\, g(T_i)^{1-\delta_i} \\ &= \Pi_{i=1}^n\, \lambda(T_i)^{\delta_i}\, \exp(-\Lambda(T_i)) \times \overline{K}(T_i)^{\delta_i}\, g(T_i)^{1-\delta_i}\,, \end{aligned}$$

where $\Lambda = -\log(1-F)$ is the cumulative hazard function, and f and g are the densities of F and K, respectively. Now, ignoring the part of the above likelihood that does not involve Λ, the log-likelihood is given by

$$l_n(\mathcal{D}, \lambda) = \sum_{i=1}^{n} \left(\delta_i \log \lambda(T_i) - \Lambda(T_i)\right). \tag{3.1}$$

We now discuss the maximum likelihood estimation procedure for the censored data model. It is not difficult to see that (3.1) cannot be meaningfully maximized over all increasing λ (since the maximum hits ∞). One way to circumvent this problem is to consider a sieved maximization scheme as employed in Marshall and Proschan (1965). However, we adopt a different route. As in Huang and Wellner (1995), we restrict the MLE of λ to be an increasing left-continuous step function with potential jumps at the $T_{(i)}$'s ($T_{(i)}$ is the i'th smallest of the T_j's and the corresponding indicator is denoted by $\delta_{(i)}$) that maximizes the right hand side of (3.1). Then, we can write,

$$\Lambda(T_{(i)}) = \sum_{j=1}^{i} \left(T_{(j)} - T_{(j-1)}\right) \lambda(T_{(j)}),$$

whence

$$l_n(D, \lambda) = \sum_{i=1}^{n} \left(\delta_{(i)} \log \lambda(T_{(i)}) - \Lambda(T_{(i)})\right) \tag{3.2}$$

$$= \sum_{i=1}^{n} \left\{ \delta_{(i)} \log \lambda(T_{(i)}) - \sum_{j=1}^{i} \left(T_{(j)} - T_{(j-1)}\right) \lambda(T_{(j)}) \right\} \tag{3.3}$$

$$= \sum_{i=1}^{n} \left\{ \delta_{(i)} \log \lambda(T_{(i)}) - (n-i+1)\left(T_{(i)} - T_{(i-1)}\right) \lambda(T_{(i)}) \right\}. \tag{3.4}$$

We then maximize (3.4) over all $0 \le \lambda_1 \le \cdots \le \lambda_n$ (where $\lambda_i \equiv \lambda(T_{(i)})$) to obtain $\hat{\lambda}_n$. This expression can indeed be meaningfully maximized. We obtain $\hat{\lambda}_n^0$, the MLE of λ under the null hypothesis $\lambda(t_0) = \theta_0$ by maximizing (3.4) over all $0 \le \lambda_1 \le \cdots \le \lambda_m \le \theta_0 \le \lambda_{m+1} \le \cdots \le \lambda_n$. Here m is such that $T_{(m)} < t_0 < T_{(m+1)}$.

The Kuhn–Tucker theorem (see, for example, Section 1.5 of Robertson, Wright and Dykstra (1988)) allows us to characterize both the unconstrained and the constrained (under the null hypothesis) MLE's of λ as solutions to isotonic regression problems. Thus we can show that $\hat{\lambda}_n(T_{(i)})$ is $\hat{f}_i$ where $\hat{f}_1 \le \cdots \le \hat{f}_n$ minimizes $\sum_{i=1}^{n} w_i(g_i - f_i)^2$ over all $0 \le f_1 \cdots \le f_n$, with

$$w_i = (n-i+1)\left(T_{(i)} - T_{(i-1)}\right) \text{ and } g_i = \frac{\delta_{(i)}}{(n-i+1)\left(T_{(i)} - T_{(i-1)}\right)}.$$

Also $\hat{\lambda}_n^0(T_{(i)})$ is $\hat{f}_i^0$ where $0 \leq \hat{f}_1^0 \leq \cdots \leq \hat{f}_n^0$ solves the constrained isotonic least squares problem: Minimize $\sum_{i=1}^n w_i(g_i - f_i)^2$ over all $0 \leq f_1 \leq \cdots \leq f_m \leq \theta_0 \leq f_{m+1} \leq \cdots \leq f_n$. For $i \neq m+1$, $\hat{\lambda}_n^0(t)$ is taken to be $\hat{\lambda}_n^0(T_{(i)})$ on $(T_{(i-1)}, T_{(i)}]$, to be θ_0 on $(T_{(m)}, t_0]$, to be $\hat{\lambda}_n^0(T_{(m+1)})$ on $(t_0, T_{(m+1)}]$.

Before proceeding further, some more notation. For points $\{(x_0, y_0), \ldots, (x_k, y_k)\}$, where $x_0 = y_0 = 0$ and $x_0 < \cdots < x_k$, consider the left-continuous function $P(x)$ such that $P(x_i) = y_i$ and such that $P(x)$ is constant on (x_{i-1}, x_i). We denote the vector of slopes (left–derivatives) of the GCM of $P(x)$ computed at the points $(x_1, \ldots, x_k)$ by $\text{slogcm}\,\{(x_i, y_i)\}_{i=0}^k$.

It is not difficult to see that $\{\hat{\lambda}_n(T_{(i)})\}_{i=1}^n = \text{slogcm}\{\sum_{j=1}^i w_j, \sum_{j=1}^i w_j\, g_j\}_{i=0}^n$, where summation over an empty set is interpreted as 0. Also, $\{\hat{\lambda}_n^0(T_{(i)})\}_{i=1}^m = \theta_0 \wedge \text{slogcm}\{\sum_{j=1}^i w_j, \sum_{j=1}^i w_j\, g_j\}_{i=0}^m$, where the minimum is interpreted as being taken componentwise, while $\{\hat{\lambda}_n^0(T_{(i)})\}_{i=m+1}^n = \theta_0 \vee \text{slogcm}\{\sum_{j=m+1}^i w_j, \sum_{j=m+1}^i w_j\, g_j\}_{i=m}^n$, where the maximum is once again interpreted as being taken componentwise.

Define the likelihood ratio statistic for testing $H_0 : \lambda(t_0) = \theta_0$ as

$$2\log \xi_n(\theta_0) = 2\left[\sum_{i=1}^n \left\{\delta_{(i)}\ \log\ \hat{\lambda}_n(T_{(i)}) - (n-i+1)\,(T_{(i)} - T_{(i-1)})\,\hat{\lambda}_n(T_{(i)})\right\}\right.$$
$$\left. - \sum_{i=1}^n \left\{\delta_{(i)}\ \log\ \hat{\lambda}_n^0(T_{(i)}) - (n-i+1)\,(T_{(i)} - T_{(i-1)})\,\hat{\lambda}_n^0(T_{(i)})\right\}\right].$$

The limit distribution of $2\log \xi_n(\theta_0)$ is established under a number of regularity conditions.

(i) Let $\tau_F = \inf\{t : F(t) = 1\}$, and let τ_H and τ_K be defined analogously. We assume that $\tau_K < \infty$ and that $0 < \tau_H = \tau_K < \tau_F$. We also assume that $0 < t_0 < \tau_H$.

(ii) F and K are absolutely continuous and their densities f and g are continuous in a neighborhood of t_0 with $f(t_0) > 0$ and $g(t_0) > 0$.

(iii) $\lambda(t)$ is continuously differentiable in a neighborhood of t_0 with $\mid \lambda'(t_0) \mid > 0$.

We are now in a position to state the key result of this paper.

Theorem 3.1. *Assume* (i), (ii) *and* (iii). *Then, under* H_0, $2\log \xi_n(\theta_0) \to_d \mathbb{D}$ *as* $n \to \infty$.

Remark. The proposed procedure has some of the flavour of empirical likelihood. Recall that, in order to maximize the likelihood function, we restricted our parameter space for λ to the class of increasing left–continuous step functions with potential jumps at the observed times. The likelihood was then maximized in this class, once under no constraints and once under the constraint that the

hazard function at a point assumes a fixed value. This is analogous, in spirit, to what is done in classical empirical likelihood. For example, for inference on the mean of a univariate distribution based on i.i.d. data, one restricts to all distributions that put mass only at the observed data points and maximizes the likelihood function in this restricted class of distributions under no constraints, and also under a hypothesis that the mean assumes a fixed value (see, for example, Owen (2001)). In classical empirical likelihood, the likelihood ratio statistic is a χ^2, as in regular parametric models, but this is no longer the case in the current situation, because of the *shape constraint* on λ. The limit distribution, $\mathbb{D}$, obtained in this case can however be viewed as an analogue of the χ^2_1 distribution that is obtained for likelihood ratio tests in regular one–parameter models. See, for example, the introduction of Banerjee (2007) for a discussion of this issue. For some insight into the form of the limiting likelihood ratio statistic, the integrated discrepancy between squared slopes of unconstrained and constrained convex minorants of Brownian motion with quadratic drift, see Wellner (2003) and, in particular, Theorem 5.1 of this paper, where it is shown that $\mathbb{D}$ is *exactly* the distribution of the likelihood ratio statistic for testing for the value of a monotone function perturbed by Brownian motion (a "white noise model"). The form of $\mathbb{D}$ falls out of the Cameron–Martin–Girsanov theorem on change of measure, and an integration by parts argument that invokes the properties of convex minorants of Brownian motion with drift.

Construction of confidence sets using the likelihood ratio statistic. Construction of confidence sets for $\lambda(t_0)$ based on Theorem 3.1 proceeds by standard inversion. Let $2\log\xi_n(\theta)$ denote the likelihood ratio statistic computed under the null hypothesis $H_{0,\theta} : \lambda(t_0) = \theta$. Then, an asymptotic level $1-\alpha$ confidence set for $\lambda(t_0)$ is given by $\{\theta : 2\log\xi_n(\theta) \le q(\mathbb{D}, 1-\alpha)\}$, where $q(\mathbb{D}, 1-\alpha)$ is the $(1-\alpha)$'th quantile of $\mathbb{D}$. Thus, finding the confidence set simply amounts to computing the likelihood ratio statistic under a family of null hypotheses. Quantiles of $\mathbb{D}$, based on discrete approximations to Brownian motion, are available in Banerjee and Wellner (2005a, Table 1).

Decreasing hazards. The result on the limit distribution of the likelihood ratio statistic in Theorem 3.1 also holds if the hazard function is decreasing. In this case, the unconstrained and constrained MLE's of λ are no longer characterized as the slopes of greatest convex minorants, but as slopes of least concave majorants. We present the characterizations of the MLE's in the decreasing hazard case below. We first introduce some notation. For points $\{(x_0, y_0), \ldots, (x_k, y_k)\}$, where $x_0 = y_0 = 0$ and $x_0 < \cdots < x_k$, consider the right–continuous function $P(x)$ such that $P(x_i) = y_i$ and such that $P(x)$ is constant on (x_{i-1}, x_i). We denote the vector of slopes (left–derivatives) of the LCM (least concave majorant) of $P(x)$ computed at the points $(x_1, \ldots, x_k)$ by $\text{slolcm}\,\{(x_i, y_i)\}_{i=0}^k$.

With $\{w_j, g_j\}_{j=1}^n$ as before, it is not difficult to see that $\{\hat{\lambda}_n(T_{(i)})\}_{i=1}^n = \text{slolcm}\{\sum_{j=1}^i w_j, \sum_{j=1}^i w_j\, g_j\}_{i=0}^n$, where summation over an empty set is interpreted as 0. Also, the MLE under $H_0 : \lambda(t_0) = \theta_0$ is given by $\{\hat{\lambda}_n^0(T_{(i)})\}_{i=1}^m = \theta_0 \vee \text{slolcm}\{\sum_{j=1}^i w_j, \sum_{j=1}^i w_j\, g_j\}_{i=0}^m$, where the maximum is interpreted as being taken componentwise, while $\{\hat{\lambda}_n^0(T_{(i)})\}_{i=m+1}^n = \theta_0 \wedge \text{slolcm}\{\sum_{j=m+1}^i w_j, \sum_{j=m+1}^i w_j\, g_j\}_{i=m}^n$, where the minimum is once again interpreted as being taken componentwise.

Unimodal hazards. Suppose now that the hazard function is unimodal. Thus there exists $M > 0$ such that the hazard function is increasing on $[0, M]$ and decreasing to the right of M, with the derivative at M being equal to 0. The goal is to construct a confidence set for the hazard function at a point $t_0 \neq M$. We consider the more realistic case for which M is unknown.

First compute a consistent estimator, $\hat{M}_n$, of the mode M. With probability tending to 1, $t_0 < \hat{M}_n$ if t_0 is to the left of M and $t_0 > \hat{M}_n$ if t_0 is to the right of M.

Assume first that $t_0 < M \wedge M_n$. Let m_n be such that $T_{(m_n)} \leq \hat{M}_n < T_{(m_n+1)}$. Let $\hat{\lambda}_n$ denote the unconstrained MLE of λ, using $\hat{M}_n$ as the mode. Then $\hat{\lambda}_n$ is obtained by maximizing (3.4) over all $\lambda_1, \ldots, \lambda_n$ with $\lambda_1 \leq \cdots \leq \lambda_{m_n}$ and $\lambda_{m_n+1} \geq \lambda_{m_n+2} \geq \ldots \geq \lambda_n$. It is not difficult to verify that $\{\hat{\lambda}_n(T_{(i)})\}_{i=1}^{m_n} = \text{slogcm}\{\sum_{j=1}^i w_j, \sum_{j=1}^i w_j g_j\}_{i=0}^{m_n}$, while $\{\hat{\lambda}_n(T_{(i)})\}_{i=m_n+1}^n = \text{slolcm}\{\sum_{j=m_n+1}^i w_j, \sum_{j=1}^i w_j\, g_j\}_{i=m_n}^n$. Now consider testing the (true) null hypothesis that $\lambda(t_0) = \theta_0$. Let $m < m_n$ be the number of $T_{(i)}$'s that do not exceed t_0. Denoting, as before, the constrained MLE by $\hat{\lambda}_n^0(t)$, it can be checked that $\hat{\lambda}_n^0(T_{(j)}) = \hat{\lambda}_n(T_{(j)})$ for $j > m_n$, whereas $\{\hat{\lambda}_n^0(T_{(i)})\}_{i=1}^m = \theta_0 \wedge \text{slogcm}\{\sum_{j=1}^i w_j, \sum_{j=1}^i w_j\, g_j\}_{i=0}^m$ and $\{\hat{\lambda}_n^0(T_{(i)})\}_{i=m+1}^{m_n} = \theta_0 \vee \text{slogcm}\{\sum_{j=m+1}^i w_j, \sum_{j=m+1}^i w_j\, g_j\}_{i=m}^{m_n}$. The likelihood ratio statistic for testing $\lambda(t_0) = \theta_0$, denoted by $2 \log \xi_n(\theta_0)$, is

$$2\Bigg[\sum_{i=1}^{m_n} \delta_{(i)}(\log\, \hat{\lambda}_n(T_{(i)}) - \log\, \hat{\lambda}_n^0(T_{(i)})) \\ - \sum_{i=1}^{m_n} (n-i+1)\,(T_{(i)} - T_{(i-1)})\,(\hat{\lambda}_n(T_{(i)}) - \hat{\lambda}_n^0(T_{(i)}))\Bigg].$$

As in the monotone hazard case, $2 \log \xi_n(\theta_0)$ converges in distribution to $\mathbb{D}$ under the assumptions of Theorem 2.1, and the asymptotic distribution of $\hat{\lambda}_n(t_0)$ is similar to that in the monotone function case. For a similar result for the maximum likelihood estimator, in the setting of unimodal density estimation away from the mode, we refer the reader to Theorem 1 of Bickel and Fan (1996). A

rigorous derivation in our problem involves some embellishments of the arguments in Section 6 of the paper and are omitted. Intuitively, it is not difficult to see why the asymptotic behavior remains unaltered. The characterization of the MLE on the interval $[0, M_n]$, with M_n converging to M is in terms of unconstrained/constrained slopes of convex minorants exactly as in the monotone function case. Furthermore, the behavior at the point t_0, which is bounded away from M_n with probability increasing to 1, is only influenced by the behavior of localized versions of the processes V_n and G_n in a shrinking $n^{-1/3}$ neighborhood of the point t_0 (where the unconstrained and the constrained MLE's differ), and these behave asymptotically in exactly the same fashion as for the monotone hazard case. Consequently, the behavior of the MLE's and the likelihood ratio statistic are unaffected. An asymptotic confidence interval of level $1-\alpha$ for $\lambda(t_0)$ can therefore be constructed as in the monotone function case.

The other situation is when $M \vee M_n < t_0$. In this case $\hat{\lambda}_n$ has the same form as above. Now, consider testing the (true) null hypothesis that $\lambda(t_0) = \theta_0$. Let m be the number of $T_{(i)}$'s such $M_n < T_i \le t_0$. Now, $\hat{\lambda}_n^0(T_{(j)}) = \hat{\lambda}_n(T_{(j)})$ for $1 \le j \le m_n$, while $\{\hat{\lambda}_n(T_{(i)})\}_{i=m_n+1}^{m_n+m} = \theta_0 \vee \text{slolcm}\{\sum_{j=m_n+1}^{i} w_j, \sum_{j=1}^{i} w_j\, g_j\}_{i=m_n}^{m_n+m}$ and $\{\hat{\lambda}_n(T_{(i)})\}_{i=m_n+m+1}^{n} = \theta_0 \wedge \text{slolcm}\{\sum_{j=m_n+m+1}^{i} w_j, \sum_{j=1}^{i} w_j\, g_j\}_{i=m_n+m}^{n}$. The likelihood ratio statistic, $2\log \xi_n(\theta_0)$, is

$$2\Bigg[\sum_{i=m_n+1}^{n} \delta_{(i)}(\log\, \hat{\lambda}_n(T_{(i)}) - \log\, \hat{\lambda}_n^0(T_{(i)})) \\ - \sum_{i=m_n+1}^{n} (n-i+1)\,(T_{(i)} - T_{(i-1)})\,(\hat{\lambda}_n(T_{(i)}) - \hat{\lambda}_n^0(T_{(i)}))\Bigg],$$

and converges in distribution to $\mathbb{D}$ as above. Confidence sets may be constructed in the usual fashion.

U–shaped hazards. Our methodology extends also to U-shaped hazards. A U-shaped hazard is a unimodal hazard turned upside down (we assume a unique minimum for the hazard). As in the unimodal hazard case, once a consistent estimator of the point at which the hazard attains its minimum has been obtained, the likelihood ratio test for the null hypothesis $\lambda(t_0) = \theta_0$ can be conducted in a manner similar to the unimodal case. The alterations of the above formulas that need to be made are quite obvious, given that the hazard is now intially decreasing and then increasing. We omit these details. The limit distribution of the likelihood ratio statistic is, of course, given by $\mathbb{D}$ (under the conditions of Theorem 2.1).

Consistent estimation of the mode. It remains to prescribe a consistent estimate of the mode in the unimodal case. Let $\hat{\lambda}^{(k)}$ be the MLE of λ based

on $\{(\Delta_{(j)}, T_{(j)}), j \neq k\}$, assuming that the mode of the hazard is at $T_{(k)}$ (so the log–likelihood function is maximized subject to λ increasing on $[0, T_{(k)}]$ and decreasing to the right of $T_{(k)}$), and let $l_{n,k}$ be the corresponding maximized value of the log–likelihood function. Then a consistent estimate of the mode is given by $T_{(k^\star)}$, where $k^\star = \text{argmax}_{1\leq k\leq n}\, l_{n,k}$. For a similar estimator in (a) the setting of a unimodal density and (b) for a unimodal regression function, see Bickel and Fan (1996) and Shoung and Zhang (2001), respectively. An analogous prescription applies to a U–shaped hazard.

4. Simulation Studies and Data Analysis

4.1.Simulation Studies

In this section, we illustrate the performance of the likelihood ratio method (LR in the tables that follow) and competing methods for constructing confidence sets for a monotone hazard rate at a point of interest, in a right–censored setting.

Simulation A. Two settings are considered. In each, the X_i's come from a Weibull distribution with $F(x) = 1 - \exp(-x^2/2)$, whence $\lambda(x) = x$, and the Y_i's follow a uniform distribution on $(0, b)$. The first setting has $b = 4$ (light censoring at 30%), and the second setting has $b = 1.5$ (heavy censoring at 70%). In each case, we are interested in estimating λ at the point $t_0 = \sqrt{2\log 2}$, the median of F. When $b = 4$, 1,500 replicates are generated for each sample size (the sequence of chosen sample sizes is displayed in Table 1); when $b = 1.5$, 6,000 replicates are generated for each chosen sample size (shown in Table 2). For each n, the average length (AL) of nominal 95% (asymptotic) C.I's for $\lambda(t_0)$ and their observed coverage (C), are recorded for each of three different methods and displayed in the corresponding table. The three different methods are (a) likelihood ratio inversion, (b) model–based parameter estimation using the limit distribution of the MLE and (c) subsampling (also using the limit distribution of the MLE).

The parameter estimation based procedure is briefly described below. From Theorem 2.2 of Huang and Wellner (1995), we have: $n^{1/3}\,(\hat\lambda_n(t_0) - \lambda(t_0)) \to_d (4\lambda(t_0)\lambda'(t_0)/\overline{H}(t_0))^{1/3}\,\mathbb{Z}$, whence, an approximate asymptotic level $1-\alpha$ confidence interval for $\lambda(t_0)$ is given by

$$\left[\hat\lambda_n(t_0) - n^{-\frac{1}{3}} q(\mathbb{Z}, 1-\frac{\alpha}{2})\Big(\frac{4\hat\lambda(t_0)\hat\lambda'(t_0)}{\widehat{\overline{H}(t_0)}}\Big)^{\frac{1}{3}},\ \hat\lambda_n(t_0) + n^{-\frac{1}{3}} q(\mathbb{Z}, 1-\frac{\alpha}{2})\Big(\frac{4\hat\lambda(t_0)\hat\lambda'(t_0)}{\widehat{\overline{H}(t_0)}}\Big)^{\frac{1}{3}}\right],$$

where $\hat\lambda(t_0), \hat\lambda'(t_0), \widehat{\overline{H}(t_0)}$ are consistent estimates of the corresponding population parameters. For $\alpha = 0.05$, $q(\mathbb{Z}, 1-\alpha/2)$ is approximately 0.99818. The method PE uses such specific estimates: $\lambda(t_0)$ is estimated by the MLE $\hat\lambda_n(t_0)$,

Table 1. Simulation setting A.1. Average length (AL) and empirical coverage (C) of asymptotic 95% confidence intervals using likelihood ratio (LR), subsampling based (SB) and parameter–estimation based (PE) methods.

	LR		SB		PE	
n	AL	C	AL	C	AL	C
50	1.283	0.927	1.570	0.939	1.565	0.955
100	0.980	0.939	1.103	0.947	1.191	0.957
200	0.767	0.943	0.933	0.970	0.917	0.947
500	0.549	0.947	0.592	0.953	0.653	0.957
1000	0.426	0.945	0.447	0.931	0.503	0.954
1500	0.372	0.940	0.392	0.942	0.434	0.953
2000	0.338	0.946	0.358	0.943	0.391	0.961
5000	0.247	0.945	0.265	0.957	0.283	0.947

$\overline{H}(t_0)$ is estimated by the empirical proportion of T_i's that exceed t_0, while $\lambda'(t_0)$ is estimated as the slope of the straight line that best fits the MLE $\hat{\lambda}$ (in the sense of least squares). While this method gives a consistent estimate of $\lambda'(t_0)$, it is not generally applicable for derivative estimation. We adopt this method for our simulation studies as it provides a computationally inexpensive way of estimating the derivative. The subsampling based method (SB) was implemented by drawing a large number of subsamples of size $b < n$ from the original sample, without replacement, and estimating the limiting quantiles of $\mid n^{1/3}(\hat{\lambda}_n(t_0) - \lambda(t_0)) \mid$ using the empirical distribution of $\mid b^{1/3}\,(\hat{\lambda}^{\star}_n(t_0) - \hat{\lambda}_n(t_0)) \mid$; here $\hat{\lambda}^{\star}_n(t_0)$ denotes the estimate of the hazard rate at the point t_0, based on the subsample. For consistent estimation of the quantiles, b/n should converge to 0 as n increases. A data driven choice of b (the block–size) is often resorted to, but can be computationally intensive; see, for example, Sections 2 and 3 of Banerjee and Wellner (2005b) for more discussion on this issue in the context of current status data (a similar model exhibiting $n^{1/3}$ asymptotics). For our simulation experiments we did not use data–driven blocksize selection. Since the data generating process is known to us, we generated separate data sets (1,000 replicates) for each sample size (and for each simulation setting), and computed subsampling based intervals for $\lambda(t_0) = t_0$ using a selection of block–sizes. We then computed the empirical coverage of the 1,000 C.I's produced for each block–size, and chose the optimal block–size for the simulations presented here, as the one for which the empirical coverage was closest to 0.95. The likelihood ratio method was implemented as described in the previous section, by inverting a family of null hypotheses $H_{0,\theta} : \lambda(t_0) = \theta$, with θ being allowed to vary on a fine grid between 0 and 6.

Table 1 shows the performances of the three methods at low censoring. The likelihood ratio based C.I's are shorter, on average, in comparison to the other

Table 2. Simulation setting A.2: Average length (AL) and empirical coverage (C) of asymptotic 95% confidence intervals using likelihood ratio (LR), subsampling based (SB) and parameter–estimation based (PE) methods.

	LR		SB		PE	
n	AL	C	AL	C	AL	C
50	3.110	0.911	9.502	0.950	2.647	0.893
100	2.408	0.917	3.270	0.970	1.972	0.906
200	1.684	0.929	2.050	0.972	1.462	0.917
500	1.073	0.932	1.283	0.974	1.016	0.927
1000	0.782	0.936	0.937	0.975	0.781	0.942
1500	0.653	0.941	0.785	0.977	0.669	0.944

methods for each displayed sample size. The likelihood ratio intervals are anticonservative at $n = 50$, but exhibit steady coverage between 94% and 95% from $n = 200$ onwards. The PE based method gives higher coverage than the LR based method and is generally conservative, which is not surprising in view of the larger intervals that it produces. The subsampling based method shows the greatest fluctuations in terms of coverage, dropping from 97% at $n = 200$ to 93% at $n = 1,000$ and rising again to around 96% at $n = 5,000$. Table 2 shows simulation results for the second setting, where we have heavy censoring. In this case, the median $\sqrt{2 \log 2} = 1.18$ (approximately) is fairly close to the right end of the support of the distribution of the observation times (1.5), and the estimation problem is more difficult than in the first setting. Consequently, the C.I's produced in this setting are wider than the corresponding C.I's in the previous case, and the small sample coverage of the LR and the PE based C.I's both suffer. However, in this case, the LR intervals are *wider* than the PE intervals at smaller sample sizes; both are anticonservative, but the PE intervals are even more so. At higher sample sizes, the two methods essentially catch up with each other in terms of length and coverage. The subsampling based intervals are, systematically, the widest of the three (the average length of 9.5 at $n = 50$ is the outcome of heavy right skewness in the length of the subsampling based C.I. and the median length, 3.34, is more reflective of the location of the distribution), and quite conservative. Furthermore, while the other methods produce observed coverage that approach the nominal, this does not happen with the subsampling based C.I's.

The above observations show that the LR method performs quite well against the two competing methods, producing C.I's that trade off coverage and length nicely. This is especially notable in light of the fact that the competition among the three methods is not completely fair, since both the PE and SB methods enjoyed the advantage of background knowledge (the manner of derivative estimation in the PE method, and the determination of optimal block size for

the subsampling method), which the likelihood ratio method did not have and more importantly, did not require. Complete data–driven estimation of nuisance parameters for the PE and SB procedures introduces more variability into the confidence intervals, and can affect the performance of these methods adversely. These issues are absent with the likelihood based procedure and make it an attractive choice.

Simulation B. For this simulation, the X_i's come from a Weibull distribution with $F(x) = 1 - \exp(-x^3/3)$, so that $\lambda(x) = x^2$, and the Y_i's follow a uniform distribution on $(0, 2)$. The goal is to estimate $\lambda(t_0) = t_0^2$, where $t_0 = (3 \log 2)^{1/3}$ (approximately 1.28) is the median of F. We study the performance of the likelihood ratio method in comparison to two kernel–based procedures: the first relies on bootstrapping (BS) using an estimate of the optimal local bandwidth, and the second is a recent procedure (CHT) advocated in Cheng, Hall and Tu (2006) that avoids bootstrapping. The average length and coverage of (asymptotic) 95% C.I's are presented for each method, based on 2,000 replicates for each sample size (here, we restrict ourselves to moderate sample sizes), and are displayed in Table 3.

To implement the bootstrap based method, we employed the Epanechnikov kernel to compute $\tilde{\lambda}(t_0)$, the usual smoothed version of the Nelson–Aalen estimator (see, for example, page 12 of Wang (2005)), using an estimate of the optimal local bandwith (display (23) in Wang (2005)) that is of order $n^{-1/5}$; this corresponds to the fact that the Epanechnikov kernel has order 2. The formula for the optimal bandwidth at a point t_0 involves the unknown quantities $\lambda(t_0), \overline{H}(t_0)$ and $\lambda^{(2)}(t_0)$. For (optimal) bandwidth estimation, a preliminary estimate of $\lambda(t_0)$ based on the "pilot" bandwidth $2\, n^{-1/5}$ was used, whereas $\overline{H}(t_0)$ was estimated by its empirical version based on the T_i's. To avoid derivative estimation, we used the true value of $\lambda^{(2)}(t_0) \equiv 2$. Approximate 95% C.I's were constructed by approximating the quantiles of the distribution of $n^{2/5}(\tilde{\lambda}(t_0) - \lambda(t_0))$ by those of $n^{2/5}(\tilde{\lambda}^\star(t_0) - \tilde{\lambda}(t_0))$, where $\tilde{\lambda}^\star(t_0)$ is the estimate based on a bootstrap sample. We used 2,000 bootstrap realizations from the sample to estimate the (0.025'th and the 0.975'th quantiles of the) bootstrap distribution.

For the CHT method, we used the bandwidth prescription given in display (2.4) of Cheng, Hall and Tu (2006), and obtained approximate 95% bias–ignored confidence intervals for $\lambda(t_0)$ using the formula in the topmost display on page 361 of their paper. The CHT intervals shrink with increasing sample size at rate $n^{-1/3}$, the same as the likelihood ratio intervals, while the bootstrap intervals shrink at rate $n^{-2/5}$. Table 3 shows the relative performance of these procedures. The LR intervals, on average, are the widest of the three, but also provide better coverage at small sample sizes, with the bootstrap based C.I's being quite

Table 3. Simulation B. Average length (AL) and empirical coverage (C) of asymptotic 95% confidence intervals using likelihood ratio (LR), bootstrap–based (BS) and the Cheng–Hall–Tu (CHT) methods.

	LR		BS		CHT	
n	AL	C	AL	C	AL	C
50	2.673	0.925	1.426	0.664	2.389	.876
100	2.037	0.932	1.441	0.802	1.841	.910
200	1.522	0.937	1.255	0.892	1.449	.930
400	1.159	0.943	0.906	0.947	1.154	.936
500	1.072	0.944	0.792	0.935	1.069	.942

markedly anticonservative. At higher sample sizes, the coverages of all three procedures are comparable.

A few words regarding the pros and cons of likelihood ratio estimation versus estimation based on standard smoothing techniques are in order. The performance of smoothing procedures often depends heavily on the choice of the tuning parameter. Thus the choice of bandwidth is important for kernel smoothing, especially at moderate sample sizes, while alternative procedures like spline based estimation of hazards (see, for example, Section 4 of Wang (2005) for a discussion and references) require judicious selection of a smoothing parameter. For example, estimation of the optimal bandwidth involves nuisance parameter estimation, like derivatives of the hazard function. While we bypassed this step for our simulation study, this is unavoidable with real–life data. The likelihood ratio procedure, as noted before, does not depend on tuning parameters and, in that sense, is a more automated procedure. Another attractive feature of likelihood ratio inversion is that it avoids estimation of nuisance parameters, which is hard to bypass with smoothing techniques. On the other hand, the likelihood ratio procedure is based on $n^{1/3}$ consistent estimates, while smoothing can produce faster rates of convergence under modest assumptions, and therefore, (typically) shorter confidence intervals. Furthermore, standard smoothing techniques do not require shape–restrictions, and are therefore more widely applicable. An extended simulation study comparing the likelihood ratio procedure to smoothing techniques employing advanced bandwidth selection techniques would be interesting and is left as a topic of future research.

4.2. Illustration on a data set

In this section, we apply the likelihood ratio method to construct confidence intervals for the risk (hazard rate) of developing schizophrenia in puberty and youth. The data come from the Jerusalem Perinatal Cohort Schizophrenia Study (JPSS) of approximately 92,000 individuals born between 1964 and 1976 to Israeli women living in Jersualem (and the adjoining rural areas). The data set

available to us is for around 88,000 of these individuals. For each individual, we know the minimum of time to diagnosis of schizophrenia and follow-up time. Denoting the age of schizophrenia development for the i'th individual by X_i and the follow-up time by Y_i, the available data is right–censored with $\Delta_i = 1(X_i \leq Y_i)$ denoting the indicator of diagnosis of schizophrenia and $T_i = X_i \wedge Y_i$ denoting the time at which the individual was removed from the study. Information on a number of potentially influential covariates (like sex, social class, paternal age at the time of the individual's birth) is also available. Out of these different covariates, paternal age is believed to play a major role, with higher paternal age associated with higher risk for developing schizophrenia at some point in life. Malaspina, Harlap, Fennig, Heiman, Nahon, Feldman and Susser (2001) demonstrate a steady increase in schizophrenia risk with advanced paternal age, a finding since replicated in subsequent studies. The rate of genetic mutation in paternal germ cells is known to increase significantly with age. Such increased mutation frequency has a strong clinical association with strong paternal age effects for multiple diseases and disorders including schizophrenia, possibly because of accumulating replication errors in spermatogonial cell lines.

For the purpose of this paper, we only take into account paternal age, the primary covariate of interest. We split the cohort into two groups, with the first (Group A) corresponding to individuals for whom the paternal age does not exceed 35 years (younger fathers), and the second (Group B) corresponding to individuals for which it does (older fathers), and analyze these two different groups separately. Precedents for subgroup analysis through stratification using a threshold value for paternal age exist in this setting, though the threshold can vary between 30 and 45 years. (Indeed, a data-driven choice of threshold is one of the interesting questions from an epidemiological standpoint. Furthemore, a better understanding of the functional dependence of schizophrenia risk on paternal age is also sought. However, full justice to such issues cannot be done within the context of this paper.) Kernel based estimates of the instantaneous hazard rate for these two sets of data indicate quite clearly that in either case the hazard risk is unimodal with a fairly sharp peak around age 20.

We estimated the (unimodal) hazard function of the age of schizophrenia diagnosis for each subgroup using the methods of Section 2. The modal value for Group A was estimated to be 19.86 years and that for Group B was estimated as 18.8 years. Asymptotically 95% confidence intervals for the hazard functions, in the two different groups, at a number of different ages were then constructed using the likelihood ratio method. Figure 1 shows two different estimates of the hazard function for Group A: the smooth estimate is obtained by kernel–smoothing the Nelson-Aalen estimator with bandwidth $34 \times n^{-1/5}$ (where n is the size of the group, approximately 65,600, and 34 is approximately the age–range), and the

step–wise estimate is computed using maximum likelihood. Note the "spiking problem" with the MLE in the vicinity of the mode – this is a consequence of the upward bias of the MLE at the mode, and is a well–known phenomenon in shape–restricted estimation. It is seen that the kernel estimate tracks the MLE well over the entire domain, with the exception of a small neighborhood around the mode, owing to the inconsistency referred to above. Pointwise confidence sets for a selection of ages are also exhibited in the figure. A pattern similar to Figure 1 is observed in Group B.

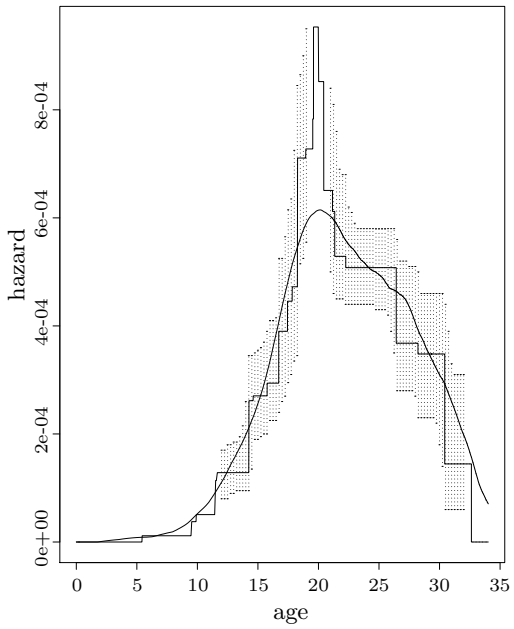

Figure 1. Estimates of the hazard function in Group A and LR based confidence sets.

Table 4 shows likelihood–ratio–based (asymptotically) 95% pointwise confidence intervals for the hazard rate at a number of selected ages for the two different groups. Because of the spiking problem, we do not report C.I's at ages 19 and 20 (these are extremely close to the estimated modes in Group B and Group A, respectively). At ages 17 and 18, the C.I's in Group B start shifting to the right of the corresponding C.I's in Group A, providing some evidence of the effect of paternal age on schizophrenia propensity. This effect is more pronounced in early youth: notice, the generally pronounced separation of the C.I's in the two groups at ages 21–24. In the late 20's the C.I.'s start overlapping once again. Since the above confidence intervals are only valid pointwise, it is important not

Table 4. 95% C.I.'s for the hazard function at selected ages in the two groups (in units of 10^{-4}).

t	C.I (A)	C.I (B)
14.00	[0.95 , 2.60]	[0.35 , 3.05]
15.00	[1.90 , 3.54]	[1.25 , 4.35]
16.00	[2.25 , 4.10]	[2.20 , 5.05]
17.00	[2.60 , 5.25]	[3.15 , 12.10]
18.00	[5.15 , 9.00]	[6.50 , 15.30]
21.00	[5.00 , 8.10]	[6.70 , 11.60]
22.00	[4.50 , 6.80]	[6.50 , 10.70]
23.00	[4.40 , 5.90]	[6.00 , 10.10]
24.00	[4.40 , 5.80]	[5.90 , 9.80]
25.00	[4.30 , 5.80]	[5.00 , 9.60]
26.00	[3.90 , 5.80]	[4.40 , 8.50]
27.00	[2.80 , 5.20]	[4.30 , 8.40]
28.00	[2.70 , 5.10]	[3.60 , 8.30]

to draw global comparisons between the hazard rates for the groups based solely on them. However, the pattern depicted in the table can be used as an initial step to identify age–ranges where differences between the two groups become more prominent, so that epidemiological features of the individuals in the sub–cohorts, defined by these age ranges, can be studied more closely.

5. Conclusion

In this paper, we have developed new methodology for pivot-based estimation of a monotone, unimodal or a U–shaped hazard, through the use of large sample likelihood ratio statistics. The most attractive feature of the proposed method is the fact that it is fully automatic and does not require estimation of nuisance parameters or smoothing parameters for its implementation. On the other hand, since the estimates underlying the likelihood ratio statistic exhibit $n^{1/3}$ rates of convergence, the procedure may not work well for very small sample sizes. In such cases, parametric fits may be more desirable from a modelling perspective.

The proposed method works for points away from the mode (in the unimodal setting) or the minimizer of the hazard (in the U shaped setting), but cannot be applied to estimation of the hazard function at the mode/the minimizer. Even if the true mode is known, naive likelihood inference for the value at the mode, which is akin to estimating a monotone function at an end–point will not work because isotonic estimators tend to be inconsistent at boundaries. The "spiking problem" in the context of estimating a monotone density at an end–point is well known. Consistent estimation at the end–point requires penalization (as in Woodroofe and Sun (1993)), or computation of the isotonic estimator at a

sequence of points converging to the end-point at an appropriate rate, with increasing sample size (Kulikov and Lopuhaa (2006)). It seems quite plausible that such techniques could be adapted to work in this situation. Yet another problem that seems to have no satisfactory solution as yet is inference for the mode of the hazard itself. While the problem of estimating the mode of a density function has been studied by a number of different authors, nonparametric large sample techniques for constructing a confidence interval for the mode, by and large, remain to be developed in the hazard setting. Shoung and Zhang (2001) derive a rate of convergence for their proposed estimator of the mode for a unimodal regression function (and an analogous result can be expected to hold in the hazard function situation), but do not derive the asymptotic distribution. A more challenging problem would be the construction of joint confidence sets for the mode and the modal value. One can envisage many different situations where this would find application, one particular instance being the schizophrenia study dealt with in Section 3.

Finally, the study of shape restricted hazard functions in semiparametric settings (as opposed to the fully nonparametric setting of this paper) also requires investigation and is expected to provide exciting avenues for future research, in particular, a more refined analysis of the data from the schizophrenia study.

Acknowledgements

The author is indebted to Dolores Malaspina and Ian McKeague at Columbia University, for making the schizophrenia data available and also for many helpful discussions, and also to Bodhi Sen for help with the simulation studies. The research was partially supported by a grant from the National Science Foundation, and by a grant from the Horace H. Rackham School of Graduate Studies, University of Michigan.

Appendix

Let $\mathbb{P}_n$ denote the empirical measure of the pair (T, δ) and let $\mathbb{Q}_n$ denote the empirical measure of the unobserved (X, Y). Take

$$V_n(t) = \int 1\{x \le y \wedge t\}\, d\,\mathbb{Q}_n(x,y) = \mathbb{P}_n\, \delta\, 1\{T \le t\} = \frac{1}{n}\sum_{i=1}^{n} \delta_i\, 1\{T_i \le t\},$$

$$\begin{aligned} G_n(t) &= \int \left((x \wedge y)\, 1\{x \wedge y \le t\} + t\, 1\{x \wedge y > t\}\right)\, d\,\mathbb{Q}_n(x,y) \\ &= \mathbb{P}_n\, \left(T\, 1\{T \le t\} + t\, 1\{T > t\}\right). \end{aligned}$$

Note that V_n is an increasing, piecewise constant, right-continuous process with a jump of $\delta_{(i)}/n$ at the point $T_{(i)}$, and that these are the only possible jumps.

On the other hand, G_n is a continuous increasing process (in t) with

$$G_n(t) = \frac{1}{n}\left(T_{(1)} + \cdots + T_{(i)} + (n-i)\,t\right) \quad , \quad t \in [T_{(i)}, T_{(i+1)})\,.$$

Note that

$$\int_{(T_{(i-1)}, T_{(i)}]} d\,G_n(t) = \frac{(n-i+1)}{n}\,(T_{(i)} - T_{(i-1)})\,. \tag{A.1}$$

Set $\xi_1(T,\delta,t) = \delta\,1\{T \le t\}$ and $\xi_0(T,\delta,t) = T\,1\{T \le t\} + t\,1\{T > t\}$. Straightforward computations show that $V(t) \equiv E(\xi_1(T,\delta,t)) = \int_0^t F(y)g(y)dy + F(t)\overline{K}(t)$, so that $V'(t) = F(t)g(t) - F(t)g(t) + f(t)\,\overline{K}(t) = f(t)\,\overline{K}(t) = \lambda(t)\,\overline{H}(t)$. Also, $G(t) \equiv E(\xi_0(T,\delta,t)) = \int_0^t x\,h(x)\,dx + t\,\overline{H}(t)$, so that $G'(t) = t\,h(t) - t\,h(t) + \overline{H}(t) = \overline{H}(t)$. It follows that $V'(t) = \lambda(t)\,G'(t)$, a fact that we will use later.

To study the likelihood ratio statistic for testing $H_0 : \lambda(t_0) = \theta_0$, we need the asymptotic behavior of the processes,

$$X_n(z) = n^{\frac{1}{3}}\left(\hat{\lambda}_n(t_0 + z\,n^{-\frac{1}{3}}) - \theta_0\right) \text{ and } Y_n(z) = n^{\frac{1}{3}}\left(\hat{\lambda}^0_n(t_0 + z\,n^{-\frac{1}{3}}) - \theta_0\right),$$

the appropriately centered and scaled versions of the MLE's of λ, treated as processes in a local time scale.

Theorem A.1. *Assume Conditions* (i)–(iii). *Take* $a = (\lambda(t_0)/\overline{H}(t_0))^{1/2} = (\theta_0/\overline{H}(t_0))^{1/2}$ *and* $b = \lambda'(t_0)/2$. *Then, under* $H_0 : \lambda(t_0) = \theta_0$,

$$(X_n(z), Y_n(z)) \to_d \left(g_{a,b}(z), g^0_{a,b}(z)\right),$$

finite–dimensionally, and also in the space $\mathcal{L} \times \mathcal{L}$, *where* $\mathcal{L}$ *is the space of functions from* $\mathbb{R} \to \mathbb{R}$ *that are bounded on every compact set, equipped with the topology of* L_2 *convergence with respect to Lebesgue measure on compact sets.*

For a proof-sketch of Theorem A.1, see Banerjee (2007).

Proof of Theorem 3.1. In what follows, we denote the set of indices i on which $\hat{\lambda}_n(T_{(i)})$ differs from $\hat{\lambda}^0_n(T_{(i)})$ by D. Let D_n denote the time interval on which $\hat{\lambda}_n$ and $\hat{\lambda}^0_n$ differ, and let $\tilde{D}_n = n^{1/3}\,(D_n - t_0)$. Now

$$\begin{aligned} 2\log\xi_n(\theta_0) = {} & 2\sum_{i=1}^{n}\delta_{(i)}\log\hat{\lambda}_n(T_{(i)}) - 2\sum_{i=1}^{n}\delta_{(i)}\log\hat{\lambda}^0_n(T_{(i)}) \\ & -2\sum_{i=1}^{n}(n-i+1)(T_{(i)} - T_{(i-1)})\,(\hat{\lambda}_n(T_{(i)}) - \hat{\lambda}^0_n(T_{(i)}))\,. \end{aligned}$$

Expanding

$$A_n \equiv 2\sum_{i=1}^{n}\delta_{(i)}\log\hat{\lambda}_n(T_{(i)}) - 2\sum_{i=1}^{n}\delta_{(i)}\log\hat{\lambda}^0_n(T_{(i)})$$

in a Taylor series around $\theta_0 \equiv \lambda(t_0)$, we get

$$A_n = 2\sum_{i\in D}\delta_{(i)}\frac{\hat\lambda_n(T_{(i)})-\theta_0}{\theta_0} - \sum_{i\in D}\delta_{(i)}\frac{(\hat\lambda_n(T_{(i)})-\theta_0)^2}{\theta_0^2}$$
$$-2\sum_{i\in D}\delta_{(i)}\frac{\hat\lambda_n^0(T_{(i)})-\theta_0}{\theta_0} + \sum_{i\in D}\delta_{(i)}\frac{(\hat\lambda_n^0(T_{(i)})-\theta_0)^2}{\theta_0^2} + r_n\,,$$

where r_n can be shown to be $o_p(1)$. Some rearrangement and rewriting of terms then yields that,

$$2\log\xi_n(\theta_0)$$
$$= \frac{2}{\theta_0}\sum_{i\in D}\Big[(\hat\lambda_n(T_{(i)})-\theta_0)-(\hat\lambda_n^0(T_{(i)})-\theta_0)\Big]\big[\delta_{(i)}-\theta_0(n-i+1)(T_{(i)}-T_{(i-1)})\big]$$
$$-\frac{1}{\theta_0^2}\sum_{i\in D}\delta_{(i)}\Big[(\hat\lambda_n(T_{(i)})-\theta_0)^2-(\hat\lambda_n^0(T_{(i)})-\theta_0)^2\Big] + o_p(1) \equiv T_1 - T_2 + o_p(1).$$

Now, consider T_1. We have

$$T_1 = \frac{2}{\theta_0}\Bigg[\sum_{i\in D}(\hat\lambda_n(T_{(i)})-\theta_0)\left(\delta_{(i)}-\theta_0(n-i+1)(T_{(i)}-T_{(i-1)})\right)$$
$$-\sum_{i\in D}(\hat\lambda_n^0(T_{(i)})-\theta_0)\left(\delta_{(i)}-\theta_0(n-i+1)(T_{(i)}-T_{(i-1)})\right)\Bigg]$$
$$= \frac{2}{\theta_0}\Bigg[\sum_{i\in D}(\hat\lambda_n(T_{(i)})-\theta_0)^2(n-i+1)(T_{(i)}-T_{(i-1)})$$
$$-\sum_{i\in D}(\hat\lambda_n^0(T_{(i)})-\theta_0)^2(n-i+1)(T_{(i)}-T_{(i-1)})\Bigg]$$
$$= \frac{2}{\theta_0}\left[\sum_{i\in D}\left((\hat\lambda_n(T_{(i)})-\theta_0)^2-(\hat\lambda_n^0(T_{(i)})-\theta_0)^2\right)(n-i+1)(T_{(i)}-T_{(i-1)})\right],$$

on using the facts that (i) D can be split up into blocks of indices, such that the constrained solution $\hat\lambda_n^0$ is constant on each block, and on any block B where the constant value c_B is different from θ_0, we have

$$c_B = \frac{\sum_{i\in B}\delta_{(i)}}{\sum_{i\in B}(n-i+1)(T_{(i)}-T_{(i-1)})}\,;$$

and (ii) the same holds true for the unconstrained solution $\hat\lambda_n$. Now, for $i \neq m+1$, $\hat\lambda_n(t) \equiv \hat\lambda_n(T_{(i)})$ for $t \in (T_{(i-1)}, T_{(i)}]$ and $\hat\lambda_n^0(t) \equiv \hat\lambda_n^0(T_{(i)})$ for $t \in (T_{(i-1)}, T_{(i)}]$. In

view of (A.1) it follows easily that

$$\Big((\hat\lambda_n(T_{(i)})-\theta_0)^2-(\hat\lambda_n^0(T_{(i)})-\theta_0)^2\Big)\,(n-i+1)\,(T_{(i)}-T_{(i-1)})$$
$$=n\int_{T_{(i-1)}}^{T_{(i)}}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dG_n(t)\,.$$

For $i=m+1$, owing to the facts that $\hat\lambda_n^0(t)$ is θ_0 for $t\in(T_{(m)},t_0]$, is $\hat\lambda_n^0(T_{(m+1)})$ for $t\in(t_0,T_{(m+1)}]$, and that these two values need not coincide, we have

$$\Big((\hat\lambda_n(T_{(m+1)})-\theta_0)^2-(\hat\lambda_n^0(T_{(m+1)})-\theta_0)^2\Big)\,(n-m)(T_{(m+1)}-T_{(m)})$$
$$=n\int_{T_{(m)}}^{T_{(m+1)}}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dG_n(t)$$
$$-n\,(\hat\lambda_n^0(T_{(m+1)})-\theta_0)^2(G_n(t_0)-G_n(T_{(m)}))\,.$$

But

$$n\,(\hat\lambda_n^0(T_{(m+1)})-\theta_0)^2\,(G_n(t_0)-G_n(T_{(m)}))$$
$$=\frac{n-m}{n}\,n^{\frac13}\,(t_0-T_{(m)})\,\Big(n^{\frac13}\,(\hat\lambda_n^0\,(T_{(m+1)})-\theta_0)\Big)^2=o_p(1)\,,$$

on using the facts that $n^{1/3}\,(T_{(m)}-t_0)$ is $o_p(1)$, that $T_{(m+1)}$ eventually lies in the difference set D_n with arbitrarily high probability, and $\sup_{t\in D_n}(n^{1/3}(\hat\lambda_n^0(t)-\theta_0))^2$ is $O_p(1)$. It follows that

$$T_1=\frac{2\,n}{\theta_0}\int_{D_n}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dG_n(t)+o_p(1)\,.$$

Also easily,

$$T_2=\frac{n}{\theta_0^2}\int_{D_n}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dV_n(t)\,.$$

Thus

$$\begin{aligned}2\log\xi_n(\theta_0)&=\frac{2\,n}{\theta_0}\int_{D_n}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dG_n(t)\\&\quad-\frac{n}{\theta_0^2}\int_{D_n}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dV_n(t)+o_p(1)\\&=\frac{2\,n}{\theta_0}\int_{D_n}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dG(t)\\&\quad-\frac{n}{\theta_0^2}\int_{D_n}\Big((\hat\lambda_n(t)-\theta_0)^2-(\hat\lambda_n^0(t)-\theta_0)^2\Big)\,dV(t)+o_p(1)\end{aligned}\tag{A.2}$$

$$
\begin{aligned}
&= \frac{2}{\theta_0} \int_{\tilde{D}_n} \left(X_n^2(z) - Y_n^2(z)\right)\, G^{'}(t_0 + n^{-\frac{1}{3}}\, z)\, dz \\
&\quad -\frac{1}{\theta_0^2} \int_{\tilde{D}_n} \left(X_n^2(z) - Y_n^2(z)\right)\, V^{'}(t_0 + n^{-\frac{1}{3}}\, z)\, dz + o_p(1) \\
&= \frac{2G^{'}(t_0)}{\theta_0} \int_{\tilde{D}_n} \left(X_n^2(z) - Y_n^2(z)\right)\, dz \\
&\quad -\frac{V^{'}(t_0)}{\theta_0^2} \int_{\tilde{D}_n} \left(X_n^2(z) - Y_n^2(z)\right) dz + o_p(1), \qquad \text{(A.3)}
\end{aligned}
$$

where (A.2) follows from the step above it on noting that

$$
\int_{D_n} \left\{ (n^{\frac{1}{3}}\, (\hat{\lambda}_n(t) - \theta_0))^2 - (n^{\frac{1}{3}}\, (\hat{\lambda}_n^0(t) - \theta_0))^2 \right\}\, d\, \left(n^{\frac{1}{3}}\, (G_n(t) - G(t)) \right)
$$

and

$$
\int_{D_n} \left\{ (n^{\frac{1}{3}}\, (\hat{\lambda}_n(t) - \theta_0))^2 - (n^{\frac{1}{3}}\, (\hat{\lambda}_n^0(t) - \theta_0))^2 \right\}\, d\, \left(n^{\frac{1}{3}}\, (V_n(t) - V(t)) \right)
$$

are $o_p(1)$, using arguments from empirical process theory. For example, the expression in the display immediately above can be rewritten as $n^{1/3}\, (\mathbb{H}_n - \mathbb{H})\, \Delta\, \Psi_n(T)$, where $\mathbb{H}_n$ is the empirical measure of the pairs $\{\Delta_i, T_i\}_{i=1}^n$, $\mathbb{H}$ denotes the joint distribution of (Δ, T) and

$$
\Psi_n(t) = \left\{ \left(n^{\frac{1}{3}}\, (\hat{\lambda}_n(t) - \theta_0) \right)^2 - \left(n^{\frac{1}{3}}\, (\hat{\lambda}_n^0(t) - \theta_0) \right)^2 \right\}\, 1_{D_n}(t)\,.
$$

But this is $o_p(1)$ on noting that the function $\Delta\, \Psi_n(T)$ eventually lies in a Donsker class of functions with arbitrarily high pre-assigned probability.

Recalling that $V^{'}(t_0) = \lambda(t_0)\, G^{'}(t_0) \equiv \theta_0\, G^{'}(t_0)$ and $G^{'}(t_0) = \overline{H}(t_0)$, from (A.3) we get

$$
\begin{aligned}
2 \log \xi_n(\theta_0) = \frac{\overline{H}(t_0)}{\theta_0} \int_{\tilde{D}_n} \left(X_n^2(z) - Y_n^2(z)\right) dz &= \frac{1}{a^2} \int_{\tilde{D}_n} \left(X_n^2(z) - Y_n^2(z)\right) dz \\
\rightarrow_d\ a^{-2} \int \left\{ (g_{a,b}(z))^2 - \left(g_{a,b}^0(z)\right)^2 \right\}\, dz\,.
\end{aligned}
$$

The last step in the above display follows from that above it by virtue of the fact that the length of $\tilde{D}_n$ is $O_p(1)$, by applying Theorem A.1 in conjunction with the Continous Mapping Theorem for distributional convergence, and the fact that $(f, g) \mapsto \int (f^2 - g^2)\, d\, \lambda$, with λ denoting Lebesgue measure, is a continuous function from $\mathcal{L} \times \mathcal{L}$ to $\mathbb{R}$. But, by Lemma 2.1,

$$
a^{-2} \int \left\{ (g_{a,b}(z))^2 - \left(g_{a,b}^0(z)\right)^2 \right\}\, dz \equiv_d \mathbb{D}\,,
$$

completing the proof.

References

Banerjee, M. (2007). Estimating monotone, unimodal and U–shaped failure rates using asymptotic pivots. Technical Report **437**, University of Michigan, Department of Statistics. Available at http://www.stat.lsa.umich.edu/~moulib/shape-failrate-pivot.pdf

Banerjee, M. (2007). Likelihood based inference for monotone response models. *Ann. Statist.* **35**, 931-956.

Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29**, 1699-1731.

Banerjee, M. and Wellner, J. A. (2005a). Score statistics for current status data: Comparisons with likelihood ratio and Wald statistics. *Internat. J. Biostatistics* **1**.

Banerjee, M. and Wellner, J. A. (2005b). Confidence intervals for current status data. *Scand. J. Statist.* **32**, 405-424.

Bickel, P. and Fan, J. (1996). Some problems on the estimation of unimodal densities. *Statist. Sinica* **6**, 23-45.

Cheng, M.-Y., Hall, P. and Tu, D. (2006). Confidence bands for hazard rates under random censorship. *Biometrika* **93**, 357-366.

Gijbels, I. and Heckman, N. E. (2004). Nonparametric testing for a monotone hazard function via normalized spacings. *J. Nonparametr. Stat.* **16**, 463-477.

Grenander, U. (1956). On the theory of mortality measurement, Part II. *Skand. Akt.* **39**, 125-153.

Groeneboom, P. and Jongbloed, G. (2003). Density estimation in the uniform deconvolution model. *Statist. Neerlandica* **57**, 136-157.

Groeneboom, P. and Wellner J. A. (2001). Computing Chernoff's distribution. *J. Comput. Graph. Statist.* **10**, 388-400.

Hall, P., Huang, L.-S., Gifford, J. A. and Gijbels, I. (2001). Nonparametric estimation of hazard rate under the constraint of monotonicity. *J. Comput. Graph. Statist.* **10**, 592-614.

Huang, J. and Wellner, J. (1995) . Estimation of a monotone density or monotone hazard under random censoring. *Scand. J. Statist.* **22**, 3-33.

Kulikov, V. N. and Lopuhaa, H. P. (2006). The behavior of the NPMLE of a decreasing density near the boundaries of the support. *Ann. Statist.* **34**, 742-768.

Malaspina, D., Harlap, S., Fennig, S., Heiman, D., Nahon, D., Feldman, D. and Susser, E. S. (2001). Advancing paternal age and the risk of schizophrenia. *Arch. Gen. Psychiatry* **58**, 361-367.

Marshall, A. W. and Proschan, F. (1965). Maximum Likelihood Estimation for distributions with monotone failure rate. *Ann. Math. Statist.* **36** , 69-77.

Mukerjee, H. and Wang, J.-L. (1993). Nonparametric maximum likelihood estimation of an increasing hazard rate for uncertain cause–of–death data. *Scand. J. Statist.* **20**, 17-33.

Mykytyn, S. W. and Santner, T. J. (1981). Maximum likelihood estimation of the survival function based on censored data under hazard rate assumptions. *Comm. Statist. Theory Methods* **A11**, 2259-2270.

Owen, A. B. (2001). *Empirical Likelihood*. Monographs on Statistics and Applied Probability **92**. Chapman and Hall.

Padgett, W. J. and Wei, L. J. (1980). Maximum likelihood estimation of a distribution function with increasing failure rate based on censored observations. *Biometrika* **67**, 470-474.

Prakasa Rao, B. L. S. (1970). Estimation for distributions with monotone failure rate. *Ann. Math. Statist.* **36** , 69-77.

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

Politis, D. M., Romano, J. P. and Wolf, M. (1999). *Subsampling*, Springer-Verlag, New York.

Shoung, J.-M. and Zhang, C.-H. (2001). Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.* **29**, 648–665.

Tsai, W. Y. (1988). Estimation of the survival function with increasing failure rate based on left truncated and right censored data. *Biometrika* **75**, 319-324.

Wang, J.-L. (1986). Asymptotically minimax estimators for distributions with increasing failure rate. *Ann. Statist.* **14**, 1113-1131.

Wang, J.-L. (2005). Smoothing hazard rate. *Encyclopedia of Biostatistics*. 2nd Edition. Vol **7**, 4986-4997. *Ann. Statist.* **28**, 779-814.

Wellner, J. A. (2003). Gaussian white noise models: some results for monotone functions. In *Crossing Boundaries: Statistical Essays in Honor of Jack Hall* **43** (Edited by J. E. Kolassa and D. Oakes), 87-104. IMS Lecture Notes–Monograph Series.

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

Department of Statistics, University of Michigan, 439 West Hall, 1085 S. University, Ann Arbor, MI 48109-1107, U.S.A.

E-mail: moulib@umich.edu

(Received March 2006; accepted November 2006)