Estimating monotone, unimodal and U–shaped failure rates using asymptotic pivots

Moulinath Banerjee

University of Michigan

March 17, 2006

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 non–standard, though pivotal, 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.

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 \( \lambda \) is the derivative of \( \Lambda \); thus, \( \lambda(x) = f(x)/(1 - F(x)) \). This quantity is called the instantaneous hazard function/failure rate, because if we consider a very small interval of time \([x, x + dx]\), then the chance that an individual fails in this time window given that they have survived up till time \( x \) is given by \( \lambda(x) dx \). A higher value of \( \lambda(x) \) therefore 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 shapes/U shapes. The 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 after a while 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 (2000) comes from industry where manufacturers often 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. If background knowledge dictates monotonicity of the hazard rate, then meaningful and efficient inference demands that such a constraint be incorporated in the statistical analysis.

Though shape constrained hazards appear quite frequently in many important applications, as noted above, there are relatively fewer nonparametric methods for estimating hazard rates, and in particular, 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 for constructing nonparametric pointwise confidence sets for a monotone failure rate that uses asymptotic pivots, and then extend this idea to the study of unimodal or U–shaped hazards. While our discussion will be framed in the realistic context of right–censored data, the proposed methodology is equally applicable to uncensored data. However, before proceeding to a discussion of our proposed method, we quickly recapitulate the setting of uncensored and right–censored data and then present a brief survey of some of the existing literature on monotone hazard estimation, with emphasis on the construction of confidence sets, to put our method and its contribution in the right perspective.

2 Background

In the uncensored case, we observe $X_1, X_2, \ldots, X_n$ which are 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 the above data. In the case of right–censored data, all the $X_i$’s are not observable; rather, one observes pairs of random variables $(T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n)$ where $\delta_i = 1\{X_i \leq Y_i\}$ and $T_i = X_i \land Y_i$ where $Y_i$ is the (random) time that the $i$’th individual is observed for. It is assumed that the $Y_i$’s are mutually independent and also 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 as before is to estimate $\lambda$.

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 $\lambda$ 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 $\lambda$ at a fixed point, was derived by Huang and Wellner (1995). This result is connected with the development in this paper, and we will return to it in more detail later. Huang and Wellner (1995) showed that the rate of convergence of $\hat{\lambda}(t_0)$, the MLE of $\lambda$ evaluated at the point $t_0$ is $n^{1/3}$; more precisely, at a fixed point $t_0$,

$$n^{1/3} \left( \hat{\lambda}(t_0) - \lambda(t_0) \right) \to C(t_0) \cdot 2Z,$$

under modest assumptions; in particular, the assumption of strict monotonicity of $\lambda$ 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 $Z = \arg\min_{h \in \mathbb{R}} (W(h) + h^2)$ with $W(h)$ being two–sided Brownian motion starting from 0. A similar result was shown to hold for a competing estimator of $\lambda$, given by the slope of the greatest convex minorant of the Nelson-Aalen estimator of the cumulative hazard function, $\Lambda$.

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: $[\hat{\lambda}(t_0) - n^{-1/3} \hat{C}(t_0) \cdot q(Z, 1 - \alpha/2), \lambda(t_0) + n^{-1/3} \hat{C}(t_0) \cdot q(Z, 1 - \alpha/2)]$, where $q(Z, 1 - \alpha/2)$ is the $(1 - \alpha/2)$'th quantile of the distribution of $Z$. Quantiles of the distribution of $Z$ are tabulated in Groeneboom and Wellner (2001). The main difficulty with the confidence set advocated above is the problem of estimating the nuisance parameter $\hat{C}(t_0)$, which among other things, depends on $\lambda'(t_0)$. Estimating the derivative of the hazard in this setting, is, however, 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 only $O_p(n^{1/3})$; consequently, direct kernel smoothing with naive bandwidth choices may not recover all the information lost in the discrete NPMLE of $\lambda$. Selection of an optimal bandwidth in terms of a 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 using the derivative of a standard kernel based estimator of the instantaneous hazard function, ignoring the monotonicity constraint (with the understanding that the sign of the estimated derivative will conform to that under the imposed monotonicity constraint, provided the sample size is reasonably large). See, for example, Sarda and Vieu (1991), Patil, Wells and Marron (1994), Gonzalez-Mantiega, Cao and Marron (1996) for access to the more recent literature on conventional kernel methods for hazard rate estimators. 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 et.al. (2001) have proposed smooth estimates of the instantaneous hazard under the constraint of monotonicity, both with uncensored and 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 as the constraint is monotone increasing/decreasing). 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 is employed in practice by discretization to a very fine grid and then resorting 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 Section 2 of their paper, 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 the scenario of unimodal/U-shaped hazards and dispenses with some of the issues that we have highlighted with 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 lie in the facts that (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 extremely inexpensive, as it requires 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 our new method does not lead to a new estimator of the instantaneous hazard (unlike the method proposed by Hall et. al. (2001)), as it bases itself on unconstrained and constrained MLE’s of $\lambda$.

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 (2005).

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 $\text{slogcm}(f, I)$ denote the left–hand slope of the GCM (greatest convex minorant) of the restriction of $f$ to the interval $I$. We abbreviate $\text{slogcm}(f, \mathbb{R})$ to $\text{slogcm}(f)$. Also define:

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

For positive constants $c$ and $d$ define the process $X_{c,d}(z) = c W(z) + dz^2$, where $W(z)$ is standard two-sided Brownian motion starting from 0. Set $g_{c,d} = \text{slogcm}(X_{c,d})$ and $g_{c,d}^0 = \text{slogcm}^0(X_{c,d})$. It is known that $g_{c,d}$ is a piecewise constant increasing function, with finitely many jumps in any compact interval. Also $g_{c,d}^0$, 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, $g_{c,d}^0$ is identically 0 in some random neighbourhood of 0, whereas $g_{c,d}$ is almost surely non-zero in some random neighbourhood of 0. Also, the length of the interval $D_{c,d}$ on which $g_{c,d}$ and $g_{c,d}^0$ differ is $O_p(1)$. For more detailed descriptions of the processes $g_{c,d}$ and $g_{c,d}^0$, see Banerjee and Wellner (2001) and Wellner (2003). Thus, $g_{1,1}$ and $g_{1,1}^0$ 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 $g_{c,d}^0$ can be related in distribution to the canonical slope processes $g_{1,1}$ and $g_{1,1}^0$. This leads to the following lemma.
For positive $a$ and $b$, set
\[ D_{a,b} = \int \left\{ (g_{a,b}(u))^2 - (g_{a,b}^0(u))^2 \right\} \, du. \]
Abbreviate $D_{1,1}$ to $D$. We have:

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


### 3 The Likelihood Ratio Method

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

**The model with right censoring:** Here we have $n$ underlying i.i.d. pairs of non-negative random variables, $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, and $X_i$ 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 they are observed for. We observe $(T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n)$ where $\delta_i = \mathbb{1}\{X_i \leq Y_i\}$ 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:

\[ H(x) = (1 - F(x))(1 - K(x)) \equiv F(x)K(x). \]

With $D \equiv ((T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n))$, the likelihood function for the data can be written as:

\[
L_n(D, \lambda) = \prod_{i=1}^{n} \left( \frac{f(T_i)}{F(T_i)} \right)^{\delta_i} \left( \frac{g(T_i)}{F(T_i)} \right)^{1-\delta_i} \phi(T_i)^{\delta_i} \exp(-\Lambda(T_i)) \times K(T_i)^{\delta_i} g(T_i)^{1-\delta_i},
\]

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 $\Lambda$ (and is therefore irrelevant as far as maximizing the likelihood with respect to $\lambda$, the derivative of $\Lambda$ is concerned), the log-likelihood is given by,

\[
l_n(D, \lambda) = \sum_{i=1}^{n} \left( \delta_i \log \lambda(T_i) - \Lambda(T_i) \right).
\]  

We now discuss the maximum likelihood estimation procedure for the censored data model. It is not difficult to see that the expression (3.1) cannot be meaningfully maximized over all increasing
\( \lambda \) (since the maximum hits \( \infty \)). 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 \( \lambda \) to be an increasing left-continuous step function with potential jumps at the \( T_{(i)} \)'s. \( T_{(i)} \) is the \( i \)th largest 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,

\[
\Lambda(T_{(i)}) = \sum_{j=1}^{i} (T_{(j)} - T_{(j-1)}) \lambda(T_{(j)}) ,
\]

whence

\[
l_n(D, \lambda) = \sum_{i=1}^{n} (\delta_{(i)} \log \lambda(T_{(i)}) - \Lambda(T_{(i)})) = \sum_{i=1}^{n} \left\{ \delta_{(i)} \log \lambda(T_{(i)}) - \sum_{j=1}^{i} (T_{(j)} - T_{(j-1)}) \lambda(T_{(j)}) \right\} = \sum_{i=1}^{n} \{ \delta_{(i)} \log \lambda(T_{(i)}) - (n - i + 1) (T_{(i)} - T_{(i-1)}) \lambda(T_{(i)}) \} .
\]

We then maximize (3.4) over all \( 0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \) (where \( \lambda_i \equiv \lambda(T_{(i)}) \)) to obtain \( \hat{\lambda}_n \). This expression can indeed be meaningfully maximized. We obtain \( \hat{\lambda}_n \), the MLE of \( \lambda \) under the null hypothesis \( \lambda(t_0) = \theta_0 \) by maximizing (3.4) over all \( 0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m \leq \theta_0 \leq \lambda_{m+1} \leq \ldots \leq \lambda_n \). Here \( m \) is such that \( T_{(m)} < t_0 < T_{(n+1)} \).

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 \( \lambda \) as solutions to isotonic regression problems. Thus, we can show that \( \hat{\lambda}_n(T_{(i)}) \) is \( \hat{f}_i \) where \( \hat{f}_1 \leq \hat{f}_2 \leq \ldots \leq \hat{f}_n \) minimizes \( \sum_{i=1}^{n} w_{i}(g_{i} - \hat{f}_{i})^2 \) over all \( 0 \leq f_1 \leq f_2 \leq \ldots \leq f_n \), with

\[
w_{i} = (n - i + 1) (T_{(i)} - T_{(i-1)}) \text{ and } g_{i} = \frac{\delta_{(i)}}{(n - i + 1) (T_{(i)} - T_{(i-1)})} .
\]

Also \( \hat{\lambda}_n^0(T_{(i)}) \) is \( \hat{f}_i^0 \) where \( 0 \leq f_1^0 \leq f_2^0 \leq \ldots \leq f_n^0 \) solves the constrained isotonic least squares problem: Minimize \( \sum_{i=1}^{n} w_{i}(g_{i} - f_{i})^2 \) over all \( 0 \leq f_1 \leq f_2 \leq \ldots \leq f_m \leq \theta_0 \leq f_{m+1} \leq \ldots \leq f_n \). For \( i \neq m + 1, \hat{\lambda}_n^0(t) \) is taken to be identically equal to \( \hat{\lambda}_n^0(T_{(i)}) \) on \( T_{(i-1)}, T_{(i)} \); on \( T_{(m)}, t_0 \], \( \hat{\lambda}_n^0(t) \) is taken to be identically equal to \( \theta_0 \) and on \( t_0, T_{(m+1)} \) it is taken to be identically equal to \( \hat{\lambda}_n^0(T_{(m+1)}) \).

Before proceeding further, some notation: For points, \( \{ (x_0, y_0), (x_1, y_1), \ldots, (x_k, y_k) \} \) where \( x_0 = y_0 = 0 \) and \( x_0 < x_1 < \ldots < x_k \), 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 will denote the vector of slopes (left–derivatives) of the GCM of \( P(x) \) computed at the points \( (x_1, x_2, \ldots, x_k) \) by \( \text{slogcm} \{ (x_i, y_i) \}_{i=0}^{k} \).
It is not difficult to see that
\[
\{\hat{\lambda}_n(T(i))\}_{i=1}^n = \text{slogcm} \left\{ \sum_{j=1}^i w_j, \sum_{j=1}^i w_j g_j \right\}_{i=0}^n,
\]
where summation over an empty set is interpreted as 0. Also,
\[
\{\hat{\lambda}_0^0(T(i))\}_{i=1}^m = \theta_0 \land \text{slogcm} \left\{ \sum_{j=1}^i w_j, \sum_{j=1}^i w_j g_j \right\}_{i=0}^m,
\]
where the minimum is interpreted as being taken componentwise, while
\[
\{\hat{\lambda}_0^0(T(i))\}_{i=m+1}^n = \theta_0 \lor \text{slogcm} \left\{ \sum_{j=m+1}^i w_j, \sum_{j=m+1}^i w_j g_j \right\}_{i=m}^n,
\]
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
\[
2 \log \xi_n(\theta_0) = 2 \left[ \sum_{i=1}^n \left\{ \delta(i) \log \hat{\lambda}_n(T(i)) - (n - i + 1)(T(i) - T(i-1)) \right\} \hat{\lambda}_n(T(i)) \right] - \sum_{i=1}^n \left\{ \delta(i) \log \hat{\lambda}_0^0(T(i)) - (n - i + 1)(T(i) - T(i-1)) \right\} \hat{\lambda}_0^0(T(i)) \right].
\]

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

(i) Let \(\tau_F = \inf\{t : F(t) = 1\}\) and let \(\tau_H\) and \(\tau_K\) 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) Both \(F\) 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 \(|\lambda'(t_0)| > 0\).

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

**Theorem 3.1** Assume that (i), (ii) and (iii) hold. Then, under \(H_0\),
\[
2 \log \xi_n(\theta_0) \to_d \mathbb{D} \text{ as } n \to \infty.
\]
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 \( \{ \theta : 2 \log \xi_n(\theta) \leq q(D, 1 - \alpha) \} \), where \( q(D, 1 - \alpha) \) is the \( 1 - \alpha \)'th quantile of \( D \). Thus, finding the confidence set simply amounts to computing the likelihood ratio statistic under a family of null hypotheses. Quantiles of \( D \), based on discrete approximations to Brownian motion, are available in Banerjee and Wellner (2005A), Table 1.

Theorem 2.2 of Huang and Wellner (1995), which can also be derived as a special case of Theorem 6.1 in Section 6, by setting \( z = 0 \), and using the relation that \( g_{1,1}(0) \equiv_d 2 \mathbb{Z} \), also leads to confidence sets for \( \lambda(t_0) \) as described in the introduction. The constant \( C(t_0) \) involved in the limit distribution is given by \( |4 \lambda(t_0) \lambda'(t_0) / H(0) |^{1/3} \) and needs to be estimated. While estimates of \( \lambda(t_0) \) and \( H(t_0) \) are readily available, estimating \( \lambda'(t_0) \) is the difficult aspect of the problem, and the issues involved therein have been detailed in the introduction, and are not regurgitated here. One way to bypass the estimation of \( C(t_0) \) is to resort to resampling techniques; indeed, the method of subsampling (Politis, Romano and Wolf, 1999) provably works, even though the usual Efron bootstrap does not. The performance of subsampling in this problem is investigated in the next section.

The case of no-censoring: It remains to discuss the case of no-censoring. The case of no censoring, i.e. where we observe the exact lifetimes of all \( n \) individuals, can be recovered from the right censoring scenario by choosing the observation times \( Y_i \)'s to be \( \infty \), whence all the \( \delta_i \)'s are 1 and all the \( T_i \)'s are \( Y_i \)'s. The maximum likelihood estimates and the likelihood ratio statistic are computed in exactly the same fashion. The likelihood ratio statistic is asymptotically pivotal as in the right-censored case, with asymptotic pivotal distribution being given by that of \( D \). Formally, the following are the regularity conditions, under which the results for the uncensored model hold.

(i') Let \( \tau_F = \inf \{ t : F(t) = 1 \} \). We assume that \( 0 < t_0 < \tau_F \).

(ii') \( F \) is absolutely continuous and its density \( f \) is continuous in a neighborhood of \( t_0 \) with \( f(t_0) > 0 \).

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

With uncensored data, we can also construct a competing asymptotic pivot based on a statistic that measures the square of the Euclidean distance between the unconstrained and constrained solution vectors. Define \( J_n = \sum_{i=1}^n (\hat{\lambda}_n(X_i) - \hat{\lambda}_n(0)(X_i))^2 \). It can be shown that if \( H_0 : \lambda(t_0) = \theta_0 \) is true, then \( J_n \) converges in distribution to \( \lambda(t_0)^2 T \) where \( T = \int (g_{1,1}(z) - g_{1,1}(z))^2 dz \). Therefore, under \( H_0, \theta_0^{-2} J_n \) is an asymptotic pivot. The quantiles of the distribution of \( T \) are also well-tabulated; see, for example, Banerjee and Wellner (2005A). Thus, an asymptotic level \( 1 - \alpha \) confidence set for \( \lambda(t_0) \) is given by \( \{ \theta : \theta^{-2} J_n \leq q(T, 1 - \alpha) \} \), where \( q(T, 1 - \alpha) \) is the \( 1 - \alpha \)'th quantile of the distribution of \( T \). Confidence sets based on \( T \) have not been pursued in the simulations.
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 \( \lambda \) 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, \( \{(x_0, y_0), (x_1, y_1), \ldots, (x_k, y_k)\} \) where \( x_0 = y_0 = 0 \) and \( x_0 < x_1 < \ldots < x_k \), consider the right–continuous function \( P(x) \) such that \( P(x_i) = y_i \) and such that \( P(x) \) is constant on \( (x_{i-1}, x_i) \).

With \( \{w_j, g_j\}_{j=1}^n \) as before, it is not difficult to see that:

\[
\{\hat{\lambda}_n(T(i))\}_{i=1}^n = \text{slocm} \left\{ \sum_{j=1}^i w_j, \sum_{j=1}^i w_j g_j \right\},
\]

where summation over an empty set is interpreted as 0. Also, the MLE under \( H_0 : \lambda(t_0) = \theta_0 \) is given by:

\[
\{\hat{\lambda}_0^n(T(i))\}_{i=1}^m = \theta_0 \lor \text{slocm} \left\{ \sum_{j=1}^i w_j, \sum_{j=1}^i w_j g_j \right\},
\]

where the maximum is interpreted as being taken componentwise, while

\[
\{\hat{\lambda}_0^n(T(i))\}_{i=m+1}^n = \theta_0 \land \text{slocm} \left\{ \sum_{j=m+1}^i w_j, \sum_{j=m+1}^i w_j g_j \right\},
\]

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 \( t_0 < M \land M_n \). 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 \( \lambda \) using \( M_n \) as the mode. Then, \( \hat{\lambda}_n \) is obtained by maximizing (3.4) over all \( \lambda_1, \lambda_2, \ldots, \lambda_n \) with \( \lambda_1 \leq \lambda_2 \ldots \leq \lambda_{m_n} \) and \( \lambda_{m_n+1} \geq \lambda_{m_n+2} \geq \ldots \geq \lambda_n \). It is not difficult to verify that

\[
\{\hat{\lambda}_n(T(i))\}_{i=1}^{m_n} = \text{slocm} \left\{ \sum_{j=1}^i w_j, \sum_{j=1}^i w_j g_j \right\}.
\]
while
\[
\{\hat{\lambda}_n(T(i))\}_{i=m_n}^{m_n+1} = \text{slogcm} \left\{ \sum_{j=m_n+1}^{i} w_j, \sum_{j=1}^{i} w_j g_j \right\}^{n}_{i=m_n}.
\]

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
\[
\{\hat{\lambda}_n^0(T(i))\}_{i=1}^{m} = \theta_0 \land \text{slogcm} \left\{ \sum_{j=1}^{i} w_j, \sum_{j=1}^{i} w_j g_j \right\}^{m}_{i=0},
\]
and
\[
\{\hat{\lambda}_n^0(T(i))\}_{i=m+1}^{m_n} = \theta_0 \lor \text{slogcm} \left\{ \sum_{j=m+1}^{i} w_j, \sum_{j=m+1}^{i} w_j g_j \right\}^{m_n}_{i=m}.
\]

The likelihood ratio statistic for testing \(\lambda(t_0) = \theta_0\) is given by:
\[
2 \log \xi_n(\theta_0) = 2 \left[ \sum_{i=1}^{m_n} \delta(i) (\log \hat{\lambda}_n(T(i)) - \log \hat{\lambda}_n^0(T(i))) - \sum_{i=1}^{m_n} (n - i + 1) (T(i) - T(i-1)) (\hat{\lambda}_n(T(i)) - \hat{\lambda}_n^0(T(i))) \right].
\]

As in the monotone hazard case, \(2 \log \xi_n(\theta_0)\) converges in distribution to \(D\) 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 stay unaffected. An asymptotic confidence interval of level \(1 - \alpha\) for \(\lambda(t_0)\) can therefore be constructed in the exact same way, as for the monotone function case.

The other situation is when \(M \lor M_n < t_0\). 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 $M_n < T_i \leq t_0$. Now, $\hat{\lambda}_n(T(j)) = \hat{\lambda}_n(T_{(j)})$ for $1 \leq j \leq m_n$, while

$$\{\hat{\lambda}_n(T(i))\}_{i=m_n+1}^{m_n+m} = \theta_0 \lor \text{sloclm} \left\{ \sum_{j=m_n+1}^{i} w_j \, \sum_{j=1}^{i} w_j g_j \right\}_{i=m_n},$$

and

$$\{\hat{\lambda}_n(T(i))\}_{i=m_n+m+1}^{n} = \theta_0 \land \text{sloclm} \left\{ \sum_{j=m_n+m+1}^{i} w_j \, \sum_{j=1}^{i} w_j g_j \right\}_{i=m_n+m}.$$

The likelihood ratio statistic is given by:

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

$$- \sum_{i=m_n+1}^{n} (n - i + 1) (T(i) - T(i-1)) (\hat{\lambda}_n(T(i)) - \hat{\lambda}_n^0(T(i))) \right],$$

and converges in distribution to $\mathbb{D}$ as above, and 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 initially decreasing and then increasing. For the sake of conciseness, we have omitted these formulas. The limit distribution of the likelihood ratio statistic is, of course, given by $\mathbb{D}$ as well (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 $\lambda$ based on $\{\Delta(j), T(j) \neq k\}$, assuming that the mode of the hazard is at $T_{(k)}$ (so the log–likelihood function is maximized subject to $\lambda$ 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 $T_{(k^*)}$, where $k^* = \arg\max_{1 \leq k \leq n} l_{n,k}$. 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) as well as two competing methods for constructing confidence sets for a monotone hazard rate at a point of interest, in a right-censored setting. The competing methods are based on the limit distribution of the MLE of $\lambda$, and use subsampling (SB in the tables) and model based parameter estimation (PE in the tables) to estimate the quantiles of the limit distribution of the MLE.

Four simulation settings are considered. In each setting the $X_i$’s come from a Weibull distribution with $F(x) = 1 - \exp(-x^2/2)$, whence $\lambda(x) = x$, and the $T_i$’s follow a uniform distribution on $(0, b)$. The first two settings correspond to $b = 4$ and $b = 3$ respectively (these are lighter censoring scenarios (30% and 40% respectively)), while the last two settings correspond to $b = 2$ and $b = 1.5$ (heavier censoring (60% and 70% respectively)). In each case, we are interested in estimating $\lambda$ at the point $t_0 = \sqrt{2 \log 2}$. This is the median of $F$. For the first two settings, the chosen sample sizes run from 50 to 5000 along the sequence displayed in Tables 1 and 2 respectively. For the last two settings we restrict ourselves to smaller sample sizes (as shown in Tables 3 and 4) only going as far as 1500. For the first two settings, 1500 replicates are generated for each sample size $n$, and asymptotically 95% confidence intervals for $\lambda(t_0)$ are constructed via the three different methods. For each $n$, the average length (AL) over the 1500 replicates and the observed coverage (C), which is the proportion of times out of 1500 that the confidence interval contains $\lambda(t_0)$, are recorded for each of the three methods and displayed in the table. The same is done for the last two settings, the only difference being that AL and C for these settings are based on 6000 replicates for each value of $n$. The goal here is to study the performance of the three methods more minutely at smaller sample sizes under heavier censoring.

The parameter estimation based procedure is briefly described below. From Theorem 2.2 of Huang and Wellner (1995), we have:

$$n^{1/3} (\hat{\lambda}_n(t_0) - \lambda(t_0)) \rightarrow_{d} \left[ \frac{4 \lambda(t_0) \lambda'(t_0)}{H(t_0)} \right]^{1/3} Z,$$

whence, an approximate asymptotic level $1 - \alpha$ confidence interval for $\lambda(t_0)$ is given by

$$[\hat{\lambda}_n(t_0) - n^{-1/3} q(Z, 1 - \alpha/2) (4 \hat{\lambda}(t_0) \hat{\lambda}'(t_0)/\widehat{H}(t_0))^{1/3}, \hat{\lambda}_n(t_0) + n^{-1/3} q(Z, 1 - \alpha/2) (4 \hat{\lambda}(t_0) \hat{\lambda}'(t_0)/\widehat{H}(t_0))^{1/3}]$$

where $\hat{\lambda}(t_0), \hat{\lambda}'(t_0), \widehat{H}(t_0)$ are consistent estimates of the corresponding population parameters. For $\alpha = 0.05$, $q(Z, 1 - \alpha/2)$ is approximately .99818. The method PE uses such specific estimates: $\lambda(t_0)$ is estimated by the MLE $\hat{\lambda}_n(t_0)$, $\widehat{H}(t_0)$ 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). This, indeed, gives a consistent estimate of $\lambda'(t_0)$, since $\lambda$ is a linear function. It should be noted that this recipe of estimating the derivative at the
point $t_0$ will not produce a consistent estimate if $\lambda$ is non-linear and is therefore not generally applicable. We adopted this method for our simulation studies as it provides a computationally inexpensive way of estimating the derivative in this situation. Since PE heavily uses background knowledge about the shape of $\lambda$, its performance as reported in the tables can be expected to be more optimistic than what one should expect in a real-life scenario, where more data-driven methods, typically, local parametric/nonparametric smoothing with some choice of a smoothing parameter, will be required. Data-driven bandwidth selection procedures for kernel smoothing could, in principle, have been employed for our simulation studies, but we chose not to adopt that route, as this would have increased computational complexity by orders of magnitude.

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 $| n^{1/3}(\hat{\lambda}_n(t_0) - \lambda(t_0)) |$, using the empirical distribution of $| b^{1/3}(\hat{\lambda}_n^*(t) - \hat{\lambda}_n(t)) |$; here $\hat{\lambda}_n^*(t)$ denotes the estimate of the hazard rate at the point $t$, based on the subsample. For consistent estimation of the quantiles, $b/n$ should converge to 0 as $n$ increases. In the literature, $b$ is referred to as the block-size, and plays a role analogous to the bandwidth $h$ in kernel based estimation. For details, see the book *Subsampling* by Politis, Romano and Wolf (1999). The choice of $b$ can effect the precision of the confidence intervals in finite samples. A data driven choice of $b$ is often resorted to but can be computationally intensive. For a discussion of subsampling in the context of a nonparametric model exhibiting cube root asymptotics as in the present case, see Sections 2 and 3 of Banerjee and Wellner (2005B). Since the issues in the present case are similar, we do not go into an exhaustive discussion here. We should, however, point out that for our simulation experiments we did not resort to data-driven blocksize selection, which, of course, is unavoidable with real life data. Since the data generating process is known to us, we generated separate data sets (1000 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 1000 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. Thus, block-size selection was done via *pilot simulations*. A natural way to extend this idea to real data sets is using the bootstrap to generate “pilot data” from the empirical measure of the observed data and choose the block size based on the bootstrapped samples. See Delgado et.al. (2003) for the details.

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 $\theta$ being allowed to vary on a fine grid between 0 and 6.

Tables 1 and 2 show the performances of the three methods for the first two simulation settings (lower censoring rates). For each setting, the likelihood ratio based C.I’s are shorter, on an average, in comparison to the other methods for each displayed sample size. For the first simulation setting, the likelihood ratio intervals are markedly anti-conservative at $n = 50$, but the coverage improves quickly as we move to higher sample sizes, with a steady coverage between 94%
Table 1: Simulation setting 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.

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

and 95% being maintained 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 with coverage dropping from 97% at \( n = 200 \) to 93% at \( n = 1000 \) and rising again to around 96% at \( n = 5000 \). For the second simulation setting, the LR method, again, exhibits consistent coverage between 94% and 95% from \( n = 100 \) onwards while the PE based intervals which are systematically larger are almost always conservative (except for \( n = 5000 \) where the coverage drops to 94.7%). The subsampling based intervals are also almost always conservative (except for \( n = 50 \)), and while they are wider than the LR based intervals, they are shorter than the PE based intervals (except for \( n = 50 \)).

Table 3 shows simulation results for Setting 3. Here the censoring rate is around 60% and as expected, all three methods suffer as a consequence, typically producing wider confidence intervals compared to the previous scenarios. Once again, the likelihood ratio method produces the sharpest C.I’s, on an average, at each sample size. The LR intervals exhibit lower coverage in comparison to the other methods, but the coverage gets close to 94% fairly quickly and continues to stay between 94% and 95% as the sample size keeps increasing. The other methods also exhibit decent coverage; the PE based C.I’s stay conservative from \( n = 200 \) onwards, while the subsampling based C.I’s fluctuate on either side of 95% without deviating drastically from the target coverage. Note the large average length of the subsampling based C.I’s at \( n = 50 \). This is the result of heavy right skewness in the distribution of the lengths of the C.I’s; the median length is 2.40 and is a better measure of central tendency in this particular case.

In Setting 4, where we have extremely heavy censoring (at more than 70%) we notice some striking changes. 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 previous settings. The C.I’s produced in this case by each
Table 2: Simulation setting 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.

<table>
<thead>
<tr>
<th>n</th>
<th>LR</th>
<th>SB</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AL</td>
<td>C</td>
<td>AL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AL</td>
</tr>
<tr>
<td>50</td>
<td>1.410</td>
<td>0.935</td>
<td>1.850</td>
</tr>
<tr>
<td>100</td>
<td>1.053</td>
<td>0.942</td>
<td>1.222</td>
</tr>
<tr>
<td>200</td>
<td>0.803</td>
<td>0.940</td>
<td>0.906</td>
</tr>
<tr>
<td>500</td>
<td>0.584</td>
<td>0.943</td>
<td>0.598</td>
</tr>
<tr>
<td>1000</td>
<td>0.454</td>
<td>0.940</td>
<td>0.495</td>
</tr>
<tr>
<td>1500</td>
<td>0.394</td>
<td>0.942</td>
<td>0.433</td>
</tr>
<tr>
<td>2000</td>
<td>0.354</td>
<td>0.944</td>
<td>0.377</td>
</tr>
<tr>
<td>5000</td>
<td>0.258</td>
<td>0.945</td>
<td>0.280</td>
</tr>
</tbody>
</table>

method are wider than the corresponding C.I’s for the previous settings, 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 behavior of the subsampling based intervals is substantially different. They 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 admirably against the competing methods. It produces C.I’s that trade off coverage and length nicely. The subsampling based method performs well in the first three settings, but not in the fourth (at least under the set of moderate sample sizes considered here) with the observed coverage staying away from the nominal. The PE based C.I’s change in character quite abruptly from being generally conservative at higher sample sizes in Setting 3 to being fairly anticonservative (especially at smaller sample sizes) in Setting 4. The disparities between the observed coverages in Settings 3 and 4 for the PE based intervals is markedly more pronounced than for the LR based intervals. In this sense, the LR based intervals seem to be somewhat stabler: they exhibit a more gradual progression across the different settings.

Since the conclusions presented here are based on a limited set of simulation studies, it is important not to overinterpret them. Nevertheless there is evidence here to demonstrate that the likelihood ratio is a competitive procedure. This is especially so, in light of the fact that the competition is not completely fair, since both competing methods enjoyed the advantage
Table 3: Simulation setting 3: 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.

<table>
<thead>
<tr>
<th>n</th>
<th>LR</th>
<th>SB</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AL</td>
<td>C</td>
<td>AL</td>
</tr>
<tr>
<td>50</td>
<td>1.857</td>
<td>0.927</td>
<td>3.03</td>
</tr>
<tr>
<td>100</td>
<td>1.344</td>
<td>0.935</td>
<td>1.548</td>
</tr>
<tr>
<td>200</td>
<td>0.990</td>
<td>0.938</td>
<td>1.986</td>
</tr>
<tr>
<td>500</td>
<td>0.693</td>
<td>0.942</td>
<td>0.738</td>
</tr>
<tr>
<td>1000</td>
<td>0.530</td>
<td>0.946</td>
<td>0.639</td>
</tr>
<tr>
<td>1500</td>
<td>0.459</td>
<td>0.945</td>
<td>0.493</td>
</tr>
</tbody>
</table>

Table 4: Simulation setting 4: 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.

<table>
<thead>
<tr>
<th>n</th>
<th>LR</th>
<th>SB</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AL</td>
<td>C</td>
<td>AL</td>
</tr>
<tr>
<td>50</td>
<td>3.110</td>
<td>0.911</td>
<td>9.502</td>
</tr>
<tr>
<td>100</td>
<td>2.408</td>
<td>0.917</td>
<td>3.270</td>
</tr>
<tr>
<td>200</td>
<td>1.684</td>
<td>0.929</td>
<td>2.050</td>
</tr>
<tr>
<td>500</td>
<td>1.073</td>
<td>0.932</td>
<td>1.283</td>
</tr>
<tr>
<td>1000</td>
<td>0.782</td>
<td>0.936</td>
<td>0.937</td>
</tr>
<tr>
<td>1500</td>
<td>0.653</td>
<td>0.941</td>
<td>0.785</td>
</tr>
</tbody>
</table>

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 competing methods could introduce more variability into the confidence intervals, and consequently affect the performance of these methods quite adversely. Such issues are absent with the likelihood based procedure; this, in conjunction with the nice balance between coverage and length as exhibited by the LR based intervals, makes the likelihood ratio procedure an attractive choice.

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 92000 individuals born between 1964 and 1976 to Israeli women living in Jerusalem (and the adjoining rural areas).
The data set available to us is for around 88000 of these individuals. For each individual, we know the minimum of time to diagnosis of schizophrenia and the 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 et. al. (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. Furthermore, 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 (using STATA) 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 the hazard functions for the two groups (A in red, B in black) to the left of the respective (estimated) modes, along with confidence intervals for some selected ages. The substantial overlap between the red and the black confidence intervals in early adolescence indicates that the effect, if any, of paternal age on the risk of schizophrenia is not pronounced in this period. The scenario however changes in later adolescence owing to a sharp increase of the estimated hazard for Group B around 17 years. From this time on, the black confidence intervals separate out from the red ones with little or no overlap between 17 and 18.5 years, providing evidence of the effect of paternal age in this period. Since these confidence intervals are only valid pointwise, the above picture should not be used to make global comparisons (over a time interval) between the two hazard functions. However, the pattern
Figure 1: Estimated hazard rates for the two different groups to the left of the respective modes and associated confidence intervals.

depicted in the picture can be used as an useful initial step to identify age ranges where differences between the two groups become prominent, so that epidemiological features of the individuals in the subcohort (defined by the age-ranges) can be analyzed more closely. Figure 2 shows the hazard functions for the two groups, with the same colour scheme as before, to the right of the respective modes, along with (asymptotically 95%) confidence intervals at a number of selected ages. In this case, the hazard functions for the two different groups stay separated to a greater degree than in the previous case: the black curve lies consistently above the red curve, and separation between the two sets of confidence intervals is also more pronounced, especially in the range 22–25 years. Notice that the black confidence intervals tend to be longer than the red ones (reflecting greater uncertainty), since the size of Group B is about a third of the size of Group A. We have refrained from giving tables with the confidence intervals for the hazard functions, as the exact numbers do not provide significantly more information in this situation.
Figure 2: Estimated hazard rates for the two different groups to the right of the respective modes and associated confidence intervals.
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, where the derivative vanishes. 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 computing 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 considerable application, one particular instance being the schizophrenia study dealt with in Section 3.

Finally, the confidence sets that we obtain in this paper are pointwise confidence sets as opposed to confidence bands. While the procedures in this paper can be extended to take care of finitely many points, there is as yet no way of constructing likelihood based simultaneous confidence bands for monotone hazard functions, and more generally confidence bands for monotone densities or regression functions. The asymptotically pivotal nature of the pointwise likelihood ratio statistic that we have considered in this paper suggests that an interesting statistic to look at, for the construction of confidence sets for the entire hazard function, might be the supremum of these pointwise likelihood ratio statistics over an interval $[a, b]$. Thus, for testing the null hypothesis $\lambda = \lambda_0$ on the interval $[a, b]$ one would consider $L_n = \sup_{t \in [a, b]} 2 \log \xi_n(\lambda_0(t_0))$, and reject the null hypothesis for large values of this quantity. The use of the supremum of pointwise likelihood ratio statistics for constructing nonparametric confidence bands has precedents in the context of
distribution functions (see, for example, Berk and Jones (1979), and Owen, 1995). Determining the limit behavior of (a potentially normalized version of) $L_n$, which would be required for setting thresholds is, however, a formidable problem in its own right, not least owing to the fact that it will require extreme value theory for correlated random variables, in addition to analytical knowledge of the tail of the distribution of $\mathbb{D}$. This last problem, in itself, can be expected to be an intricate exercise, involving sophisticated machinery from probability theory. In fact, very little is analytically known about $\mathbb{D}$ at this point.

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, and in particular a more refined analysis of the data from the schizophrenia study.

6 Technical Derivations

Let $\mathbb{P}_n$ denote the empirical measure of the pair $(T, \delta)$ and let $\mathbb{Q}_n$ denote the empirical measure of the unobserved $(X, Y)$; we define the processes $V_n$ and $G_n$ as,

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

and

$$G_n(t) = \int ((x \wedge y) \mathbb{1}\{x \wedge y \leq t\} + t \mathbb{1}\{x \wedge y > t\}) \, d\mathbb{Q}_n(x, y) = \mathbb{P}_n (T \mathbb{1}\{T \leq t\} + t \mathbb{1}\{T > t\}).$$

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 these are the only possible jumps. On the other hand $G_n$ is a continuous increasing process (in $t$) with

$$G_n(t) = \frac{1}{n} \left( T(1) + T(2) + \ldots + T(i) + (n - i) t \right) \quad t \in [T(i), T(i+1)).$$

Note that,

$$\int_{(T(i-1), T(i))} dG_n(t) = G_n(T(i)) - G_n(T(i-1)) = \frac{(n-i+1)}{n} (T(i) - T(i-1)). \quad (6.11)$$

Set:

$$\xi_1(T, \delta, t) = \delta \mathbb{1}\{T \leq t\} \quad \text{and} \quad \xi_0(T, \delta, t) = T \mathbb{1}\{T \leq t\} + t \mathbb{1}\{T > t\}.$$

We compute, $V(t) = E(\xi_1(T, \delta, t))$ and $G(t) = E(\xi_0(T, \delta, t))$. We have,

$$V(t) = E(\delta \mathbb{1}\{T \leq t\}).$$
\[
E \left(1 \{X \leq Y \wedge t\}\right) = \int P( X \leq Y \wedge t \mid Y = y) \ g(y) \ dy = \int_0^t F(y) \ g(y) \ dy + F(t) \overline{K}(t).
\]

Thus,
\[
V'(t) = F(t)g(t) - F(t)g(t) + f(t)\overline{K}(t) = f(t)\overline{K}(t) = \lambda(t)\overline{\Pi}(t).
\]

Also,
\[
G(t) = E \left( T \ 1 \{T \leq t\} + t \ 1 \{T > t\} \right) = \int_0^t x \ h(x) \ dx + \int_t^\infty t \ h(x) \ dx = \int_0^t x \ h(x) \ dx + t \overline{H}(t),
\]

where \(h\) denotes the derivative of \(H\). Thus,
\[
G'(t) = t \ h(t) - t \ h(t) + \overline{H}(t) = \overline{H}(t).
\]

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,
\[
X_n(z) = n^{1/3} \left( \hat{\lambda}_n(t_0 + z n^{-1/3}) - \theta_0 \right) \quad \text{and} \quad Y_n(z) = n^{1/3} \left( \hat{\lambda}_0(t_0 + z n^{-1/3}) - \theta_0 \right),
\]

the appropriately centered and scaled versions of the MLE’s of \(\lambda\), treated as processes in a local time scale.

**Theorem 6.1** Assume that Conditions (i) – (iii) hold. Define:
\[
a = \sqrt{\frac{\lambda(t_0)}{\overline{H}(t_0)}} = \sqrt{\frac{\theta_0}{\overline{H}(t_0)}} \quad \text{and} \quad b = \frac{1}{2} \lambda'(t_0).
\]

Then, under \(H_0: \lambda(t_0) = \theta_0\),
\[
(X_n(z), Y_n(z)) \rightarrow_d \left( g_{a,b}(z), g^0_{a,b}(z) \right),
\]

finite–dimensionally, and also in the space \(\mathcal{L} \times \mathcal{L}\), where \(\mathcal{L}\) is the space of functions from \(\mathbb{R} \rightarrow \mathbb{R}\) that are bounded on every compact set, equipped with the topology of \(L_2\) convergence with respect to Lebesgue measure on compact sets.
**Proof–sketch of Theorem 6.1:** Define the process \( \tilde{G}_n(t) = G_n(T_{i(j)}) \) for \( t \in [T_{i(j)}, T_{i(j+1)}] \). Then \( \tilde{G}_n(t) \) is a piecewise constant right–continuous function, like \( V_n(t) \). Next, define the process

\[
\tilde{M}_n(z) = \frac{1}{\theta_0(t_0)} n^{2/3} \left[ (V_n(t_0 + z n^{-1/3}) - V_n(t_0)) - \theta_0 (\tilde{G}_n(t_0 + z n^{-1/3}) - \tilde{G}_n(t_0)) \right],
\]

and the process

\[
\tilde{U}_n(z) = n^{1/3} G'(t_0)^{-1} (\tilde{G}_n(t_0 + z n^{-1/3}) - \tilde{G}_n(t_0)).
\]

It follows from equation (3.5) and the definitions of the processes \( V_n \) and \( G_n \) that

\[
X_n(z) = \text{slogcm} \left( \tilde{M}_n \circ \tilde{U}_n^{-1} \right)(\tilde{U}_n(z)).
\]

(6.13)

Here \( \tilde{U}_n^{-1} \) is defined in the usual way for a right–continuous non–decreasing function. Also, from equations (3.6) and (3.7) it follows that

\[
Y_n(z) = \text{slogcm}^0 \left( \tilde{M}_n \circ \tilde{U}_n^{-1} \right)(\tilde{U}_n(z)).
\]

(6.14)

Define \( M_n(z) \) and \( U_n(z) \) in the same way as \( \tilde{M}_n \) and \( \tilde{U}_n \) respectively, but with \( \tilde{G}_n \) replaced by \( G_n \). It is not difficult to show that \( M_n - \tilde{M}_n \) converges to 0 in probability, uniformly on every compact set, which implies that \( U_n(z) - \tilde{U}_n(z) \) also converges to 0 in probability, uniformly on every compact set. Furthermore, under conditions (i) – (iii), \( M_n(z) \) converges in distribution to \( X_{a,b}(z) \), with \( a \) and \( b \) as defined in the statement of Theorem 6.1, under the topology of uniform convergence on compacta, while \( U_n(z) \) converges in probability to the deterministic function \( z \), uniformly on every compact set. The convergence of the process \( M_n(z) \) to its limit is established in Huang and Wellner (1995) (see Theorem 5.1 of that paper). The proof in Huang and Wellner (1995) uses strong approximation results; an alternative proof can be based on Theorems 2.11.22 and Theorems 2.11.23 of van der Vaart and Wellner (1996) which deal with central limit theorems for processes defined on classes of functions that change with \( n \), the sample size. The convergence of \( U_n(z) \) to its limit follows from elementary empirical process arguments. Conclude that \( M_n \) converges in distribution to \( X_{a,b}(z) \) and \( U_n(z) \) converges in probability to \( z \). Using continuous–mapping for slopes of greatest convex minorants, we conclude that the limit distributions of \( X_n \) and \( Y_n \) are obtained by replacing the processes on the right sides of (6.13) and (6.14) respectively, by their limits. For details of the technicalities involved in employing the continuous-mapping argument for slopes of convex minorants, in a class of related models, see the proof of Theorem 2.1 in Banerjee (2005). Thus, for any \( (z_1, \ldots, z_k) \)

\[
\{X_n(z_i), Y_n(z_i)\}_{i=1}^k \to \{\text{slogcm} X_{a,b}(z_i), \text{slogcm}^0 X_{a,b}(z_i)\}_{i=1}^k
\]

in distribution. But the vector on the right side of the above display is precisely \( \{g_{a,b}(z_i), g_{a,b}^0(z_i)\}_{i=1}^k \).

The above finite dimensional convergence, coupled with the monotonicity of the functions involved, allows us to conclude that convergence happens in the space \( \mathcal{L} \times \mathcal{L} \). The strengthening of
finite dimensional convergence to convergence in the $L_2$ metric is deduced from the monotonicity of the processes $X_n$ and $Y_n$, as in Corollary 2 of Theorem 3 in Huang and Zhang (1994). If $\phi_n, \phi$ are monotone functions such that $\phi_n$ converges to $\phi$ pointwise, then $\phi_n$ also converges to $\phi$ in the $L_2$ sense, on every compact set. \hfill \Box

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

$$2 \log \xi_n(\theta_0) = 2 \sum_{i=1}^n \delta(i) \log \hat{\lambda}_n(T(i)) - 2 \sum_{i=1}^n \delta(i) \log \hat{\lambda}_n^0(T(i)) - 2 \sum_{i=1}^n (n-i+1)(T(i)-T(i-1)) (\hat{\lambda}_n(T(i)) - \hat{\lambda}_n^0(T(i))).$$

Expanding

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

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,

$$2 \log \xi_n(\theta_0) = 2 \frac{\theta_0}{\theta_0} \sum_{i \in D} \left[ (\hat{\lambda}_n(T(i)) - \theta_0) - (\hat{\lambda}_n^0(T(i)) - \theta_0) \right] \left[ \delta(i) - \theta_0 (n-i+1) (T(i) - T(i-1)) \right]
- \frac{1}{\theta_0} \sum_{i \in D} \delta(i) \left[ (\hat{\lambda}_n(T(i)) - \theta_0)^2 - (\hat{\lambda}_n^0(T(i)) - \theta_0)^2 \right] + o_p(1) \equiv T_1 - T_2 + o_p(1).$$

Now, consider $T_1$. We have,

$$T_1 = \frac{2}{\theta_0} \left[ \sum_{i \in D} (\hat{\lambda}_n(T(i)) - \theta_0) (\delta(i) - \theta_0 (n-i+1) (T(i) - T(i-1)))
- \sum_{i \in D} (\hat{\lambda}_n^0(T(i)) - \theta_0) (\delta(i) - \theta_0 (n-i+1) (T(i) - T(i-1))) \right]
= \frac{2}{\theta_0} \left[ \sum_{i \in D} (\hat{\lambda}_n(T(i)) - \theta_0)^2 (n-i+1) (T(i) - T(i-1))
- \sum_{i \in D} (\hat{\lambda}_n^0(T(i)) - \theta_0)^2 (n-i+1) (T(i) - T(i-1)) \right]
= \frac{2}{\theta_0} \left[ \sum_{i \in D} \left( (\hat{\lambda}_n(T(i)) - \theta_0)^2 - (\hat{\lambda}_n^0(T(i)) - \theta_0)^2 \right) (n-i+1) (T(i) - T(i-1)) \right].$$
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 $\theta_0$, 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 $\hat{\lambda}_n(t) \equiv \hat{\lambda}_n^0(T(i))$ for $t \in (T(i-1), T(i)]$. In view of (6.12) it follows easily that

$$\left( (\hat{\lambda}_n(T(i)) - \theta_0)^2 - (\hat{\lambda}_n^0(T(i)) - \theta_0)^2 \right) (n - i + 1) (T(i) - T(i-1))$$

equals

$$n \int_{T(i-1)}^{T(i)} \left( (\hat{\lambda}_n(t) - \theta_0)^2 - (\hat{\lambda}_n^0(t) - \theta_0)^2 \right) dG_n(t).$$

For $i = m + 1$, owing to the facts that $\hat{\lambda}_n^0(t)$ is identically equal to $\theta_0$ for $t \in (T(m), t_0]$ and equal to $\hat{\lambda}_n^0(T(m+1))$ for $t \in (t_0, T(m+1)]$ and that these two values need not necessarily coincide, we have,

$$\left( (\hat{\lambda}_n(T(m+1)) - \theta_0)^2 - (\hat{\lambda}_n^0(T(m+1)) - \theta_0)^2 \right) (n - m) (T(m+1) - T(m))$$

equals

$$n \int_{T(m)}^{T(m+1)} \left( (\hat{\lambda}_n(t) - \theta_0)^2 - (\hat{\lambda}_n^0(t) - \theta_0)^2 \right) dG_n(t) - n (\hat{\lambda}_n^0(T(m+1)) - \theta_0)^2 (G_n(t_0) - G_n(T(m))).$$

But,

$$n (\hat{\lambda}_n^0(T(m+1)) - \theta_0)^2 (G_n(t_0) - G_n(T(m))) = \frac{n - m}{n} n^{1/3} (t_0 - T(m)) \left( n^{1/3} (\hat{\lambda}_n^0(T(m+1)) - \theta_0) \right)^2 = o_p(1),$$

on using the facts that $n^{1/3} (T(m) - t_0)$ is $o_p(1)$ and 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

$$T_1 = \frac{2n}{\theta_0} \int_{D_n} \left( (\hat{\lambda}_n(t) - \theta_0)^2 - (\hat{\lambda}_n^0(t) - \theta_0)^2 \right) dG_n(t) + o_p(1).$$

Also easily,

$$T_2 = \frac{n}{\theta_0^2} \int_{D_n} \left( (\hat{\lambda}_n(t) - \theta_0)^2 - (\hat{\lambda}_n^0(t) - \theta_0)^2 \right) dV_n(t).$$

Thus,

$$2 \log \xi_n(\theta_0) = \frac{2n}{\theta_0} \int_{D_n} \left( (\hat{\lambda}_n(t) - \theta_0)^2 - (\hat{\lambda}_n^0(t) - \theta_0)^2 \right) dG_n(t)$$

$$- \frac{n}{\theta_0^2} \int_{D_n} \left( (\hat{\lambda}_n(t) - \theta_0)^2 - (\hat{\lambda}_n^0(t) - \theta_0)^2 \right) dV_n(t) + o_p(1).$$
where (6.15) follows from the step above it on noting that

$H$ from (6.16) we get,

![Mathematical expression]

Recalling that,

with arbitrarily high pre-assigned probability.

But this is immediately above can be rewritten as,

are $o_p(1)$, using arguments from empirical process theory. For example, the expression in the display 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 $\Delta, T$ and

$\Psi_n(t) = \left\{ n^{1/3} (\hat{\lambda}_n(t) - \theta_0) - n^{1/3} (\hat{\lambda}_0(t) - \theta_0) \right\}^2 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 (6.16) we get,

$2 \log \xi_n(\theta_0) = \frac{\overline{H}(t_0)}{\theta_0} \int_{\tilde{D}_n} \left( X_n^2(z) - Y_n^2(z) \right) dz$

$= \frac{1}{a^2} \int_{\tilde{D}_n} \left( X_n^2(z) - Y_n^2(z) \right) dz$

$\rightarrow_d a^{-2} \int \left\{ (g_{a,b}(z))^2 - (g_{a,b}^0(z))^2 \right\} dz .
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 $O_p(1)$, and by applying Theorem 6.1 in conjunction with the continuous mapping theorem for distributional convergence and the fact that $(f, g) \mapsto \int (f^2 - g^2) \, d\lambda$, with $\lambda$ denoting Lebesgue measure, is a continuous function from $\mathcal{L} \times \mathcal{L}$ to $\mathbb{R}$. But, by Lemma 2.1,

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

completing the proof. $\square$

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. 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.

References


