NONPARAMETRIC ESTIMATION OF HAZARD RATE UNDER THE CONSTRAINT OF MONOTONICITY

James A. Gifford[1] Irène Gijbels[1,2]

Peter Hall[1] Li-Shan Huang[1]

ABSTRACT. We show how to smoothly 'monotonise' standard kernel estimators of hazard rate, using bootstrap weights. Our method takes a variety of forms, depending on choice of kernel estimator and on the distance function used to define a certain constrained optimisation problem. We confine attention to a particularly simple kernel approach, and explore a range of distance functions. It is straightforward to reduce 'quadratic' inequality constraints to 'linear' equality constraints, and so our method may be implemented using little more than conventional Newton-Raphson iteration. Thus, the necessary computational techniques are very familiar to statisticians. We show both numerically and theoretically that monotonicity, in either direction, can generally be imposed on a kernel hazard rate estimator regardless of the monotonicity or otherwise of the true hazard rate. Our methods have straightforward extension to the problem of testing for monotonicity of hazard rate, where the distance function plays the role of a test statistic.

KEYWORDS. Bandwidth, biased bootstrap, censored data, decreasing hazard rate, increasing hazard rate, power divergence, survival analysis.

SHORT TITLE. Hazard rate estimation.

---

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2] Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

1. INTRODUCTION

Estimation of hazard rate under monotonicity assumptions is important in survival analysis, for example as a means of introducing additional information in problems involving unidentifiable cause-of-death data (Mukerjee and Wang, 1993). The hazard rate is constant only for an exponential distribution, and cases of increasing or decreasing hazard rate are, broadly speaking, those where the tail of the sampled distribution is lighter or heavier, respectively, than that of the exponential. But despite the importance of these principles, and the potential applications, there are few nonparametric methods for estimating hazard rate under monotonicity constraints.

One approach would be via spline smoothing, for example using log-spline methods developed by Kooperberg and Stone (1991, 1992) and Kooperberg, Stone and Truong (1995a). These ideas have been used to construct unconstrained log-hazard rate estimates; see Kooperberg, Stone and Truong (1995b,c). They have potential application to hazard rate estimation under the constraint of monotonicity. See Villalobos and Wahba (1987), for example, for related applications of inequality-constrained smoothing splines. The technique is not straightforward to implement in the hazard-rate case, however.

In the spirit of Grenander (1956), and also of the Kaplan-Meier (1958) estimator, Mukerjee and Wang (1993) took a 'nonparametric maximum likelihood' approach to this constrained estimation problem. The resulting monotone hazard rate estimator is relatively rough, however, and requires smoothing if it is to be easily interpreted. From this viewpoint, hazard rate estimators based on more conventional kernel methods are particularly attractive; see the reviews by Singpurwalla and Wong (1983) and Hassani, Sarda and Vieu (1986) for an account of relatively early developments, and Müller and Wang (1990, 1994), Sarda and Vieu (1991), Patil (1993a,b), Patil, Wells and Marron (1994) and González-Manteiga, Cao and Marron (1996) for access to the more recent literature. None of the estimators that arise in this way is guaranteed to be monotone, however.

In the present paper we suggest methods for overcoming this difficulty. We take a particularly simple kernel technique for estimating hazard rate, and show how to monotonise it using a 'biased bootstrap' argument (Hall and Presnell, 1999). This involves giving weights to data values, and selecting the weights so as to minimise distance to the uniform-weight case, subject to achieving monotonicity.

The same ideas may be applied to monotonise other kernel methods for hazard rate estimation. But an attractive feature of the approach that we suggest is that it not only produces a monotone, nonparametric estimator of the hazard rate, but also gives explicit, nonparametric estimators of the density f and risk $\bar{F} = 1 - F$ (where F is the distribution function corresponding to f), which are such that their ratio, the hazard rate estimator, is monotone. Moreover, the estimator of f is itself a proper probability density, and the estimator of F is obtained by integrating the estimator of f.

Our constrained estimator may be used as the basis for a test of monotonicity on a given interval. We suggest taking the test statistic to be the minimum distance from the uniform-weight case that we must move in order to successfully impose the constraint. The test may be calibrated under the assumption that the sampled distribution is exponential on the interval; this is in a sense the most difficult of the null distributions for which to test. Double bootstrap methods may be employed to enhance level accuracy. Such approaches to calibration give a consistent test which is conservative for strictly monotone null distributions.

2. METHODOLOGY

2.1. Constrained weighted-bootstrap methods. Let $X_1, \ldots, X_n$ denote a random sample from the distribution with density f, and let F be the corresponding distribution function. The associated hazard rate is $\lambda = f/(1-F)$. Given a distribution $p = (p_1, \ldots, p_n)$ on the integers $1, \ldots, n$, let $D(p)$ denote a measure of the distance, or divergence, between p and the uniform distribution, $p_{\text{unif}} = (1/n, \ldots, 1/n)$. Put

$$\hat{f}(x|p) = h^{-1} \sum_{i=1}^{n} p_i \, K\Big(\frac{x - X_i}{h}\Big) \,,$$

$$\widehat{F}(x|p) = \int_{-\infty}^{x} \hat{f}(u|p) \, du \,, \quad \hat{\lambda}(x|p) = \frac{\hat{f}(x|p)}{1 - \widehat{F}(x|p)} \,.$$

To achieve an estimator of λ which is monotone increasing on an interval $\mathcal{I}$, or monotone decreasing there, we suggest choosing $p = \hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ to minimise $D(p)$ subject to $\hat{\lambda}'(\cdot|p) \geq 0$ or $\hat{\lambda}'(\cdot|p) \leq 0$, respectively, on $\mathcal{I}$, where $\hat{\lambda}'(x|p) = (\partial/\partial x) \, \hat{\lambda}(x|p)$. Then, our estimator of λ is $\hat{\lambda}(\cdot|\hat{p})$. Note that $\hat{\lambda}'(x|p)/\hat{\lambda}_1(x|p)$ is always nonnegative, where

$$\hat{\lambda}_1(x|p) = \{1 - \widehat{F}(x|p)\} \, \hat{f}'(x|p) + \hat{f}(x|p)^2 \,.$$

Therefore, we impose the constraint on $\hat{\lambda}_1(\cdot|p)$ rather than $\hat{\lambda}'(\cdot|p)$.

In the formulae above, K is a kernel function and h a bandwidth. In practice K would be a probability density, implying that $\int \hat{f}(\cdot|p) = 1$ for each p. There is no difficulty in formulating our method and discussing its properties if h depends on location, x, but for the sake of simplicity we have chosen not to do this. If h were a function of x then it would no longer be true that $\hat{f}$ would always integrate to 1, although the normalisation factor would not depend on p.

Our definition of $\hat{p}$ may result in $\hat{\lambda}'(\cdot|\hat{p})$ having a zero somewhere in $\mathcal{I}$. If this is not desirable then we can, instead of asking that $\pm\hat{\lambda}'(\cdot|p) \geq 0$, impose the constraint $\pm\hat{\lambda}'(\cdot|p) \geq \epsilon$, where ϵ denotes a small positive quantity.

Power-divergence measures of distance, introduced by Cressie and Read (1984), are defined by

$$D_\rho(p) = \frac{1}{\rho\,(1-\rho)} \left\{ n - \sum_{i=1}^{n} (np_i)^\rho \right\}$$

when $-\infty < \rho < \infty$ and $\rho \neq 0, 1$, and by taking limits when $\rho = 0, 1$:

$$D_0(p) = -\sum_{i=1}^{n} \log(np_i)\,, \quad D_1(p) = \sum_{i=1}^{n} p_i \, \log(np_i)\,.$$

The case $\rho = \frac{1}{2}$ corresponds to Hellinger distance. The distance D_0 was used by Owen (1988, 1990) in his development of empirical likelihood methods. For the most part we shall confine attention to $D = D_\rho$, where $0 \leq \rho \leq 1$. However, observe that for all ρ, $D_\rho(p) \geq 0$ and $D_\rho(p) = 0$ if and only if $p = p_{\text{unif}}$.

2.2. Existence of $\hat{p}$. Almost no matter what the sampled distribution and choice of $\mathcal{I}$, samples can arise for which it is impossible to select p so that $\hat{\lambda}'(\cdot|p)$ has a given sign on $\mathcal{I}$. For example, suppose K is a differentiable density function with support equal to the interval $[-1, 1]$, and h is strictly less than half the width of $\mathcal{I}$. Let $\mathcal{E}$ denote the event that at least one datum X_i lies in $\mathcal{I}$ and is distant at least h from all other data and from the ends of $\mathcal{I}$. Provided f is continuous and strictly positive on $\mathcal{I}$, the probability of this event is strictly positive. And if $\mathcal{E}$ holds, it is impossible to choose p so that $\hat{\lambda}'(\cdot|p)$ does not vanish on $\mathcal{I}$; in fact, $\hat{\lambda}'(\cdot|p)$ will change sign on $\mathcal{I}$, from strictly positive to strictly negative or vice versa, unless the p_i that corresponds to the aforementioned X_i is set equal to 0.

However, datasets which conform to the event $\mathcal{E}$, or to an event like it, are pathological in nature and are not suggestive of monotone hazard rates. If we are employing a bandwidth h that is appropriate for estimating f on $\mathcal{I}$, and if there does not exist a probability distribution p such that $\hat{\lambda}(\cdot|p)$ is monotone increasing (or decreasing) on $\mathcal{I}$, then it would be unlikely that the true λ was monotone increasing (or decreasing, respectively). Therefore, the fact that a constrained monotone estimator $\hat{f}(\cdot|\hat{p})$ does not always exist is not really a problem, not least because nonexistence provides information about the true hazard rate. See section 2.3 for discussion of the relationship between our estimation methods and techniques for testing the hypothesis of monotonicity.

In support of these arguments, Theorem 4.1 in section 4 will show that under mild regularity conditions, the probability that the constrained monotone estimator $\hat{\lambda}(\cdot|\hat{p})$ fails to exist is exponentially small as a function of sample size. The conditions do not include monotonicity of the true hazard rate λ, and the monotonicity imposed on $\hat{\lambda}(\cdot|\hat{p})$ may be either increasingness or decreasingness, regardless of λ.

2.3. Hypothesis testing. One method, based on the "most difficult" parametric alternative hypothesis, will be suggested below. It is calibrated in the same way as techniques that have been proposed for testing unimodality in density estimation (see e.g. Hartigan and Hartigan, 1985) and monotonicity in nonparametric regression (e.g. Bowman, Jones and Gijbels, 1998). However, an alternative, nonparametric technique, which we found to be less conservative in terms of coverage accuracy, will be proposed in section 3.3.

Consider testing the hypothesis H_0 that λ is nondecreasing, say, on an interval $\mathcal{I}_0 = [a, b]$, against the complementary alternative. The most difficult null distribution against which to test is one for which λ is constant on $\mathcal{I}_0$ — that is, $f \equiv f_0(\cdot|\theta)$, where $f_0(x|\theta) \equiv \exp(\theta_1 x + \theta_2)$ on $\mathcal{I}_0$, for a bivariate parameter $\theta = (\theta_1, \theta_2)$. Therefore, the test suggested below will involve calibrating against this density. Our approach is to estimate λ subject to the constraint that it is nondecreasing, using methods suggested in section 2.1, and reject H_0 if the observed value of $D(\hat{p})$ is too large.

Monte Carlo simulation may be used to determine a critical point $\hat{d}$ for this test, as follows. Assume the support of K is contained in the interval $[-1, 1]$ and that the density f is smooth on $\mathcal{I}_1 = [a - h, b + h]$. Condition on that part of the sample $\{X_1, \ldots, X_n\}$ which lies in $\mathcal{I}_1$. Suppose there are N data there, comprising the subsample $\mathcal{Y} = \{Y_1, \ldots, Y_N\}$. From $\mathcal{Y}$, using maximum likelihood, compute an estimator $\hat{\theta}$ of θ under the model $f_0(\cdot|\theta)$ for f on $\mathcal{I}_0$. Conditional on $\mathcal{Y}$, simulate a random sample $\mathcal{Y}^* = \{Y_1^*, \ldots, Y_N^*\}$ from the distribution with density $f_0(\cdot|\hat{\theta})$. (Note that $\mathcal{Y}^*$ is not drawn by resampling from $\mathcal{Y}$.) Given a bandwidth h, let $\hat{\lambda}(\cdot|p)$ and $\hat{\lambda}^*(\cdot|p)$ be analogues of the kernel estimators introduced in section 2.1, for estimation on $\mathcal{I}_0$ and computed from $\mathcal{Y}$ and $\mathcal{Y}^*$, respectively. Let $\hat{p}$ and $\hat{p}^*$ be the probability vectors chosen to minimise $D(p)$ subject to $\hat{\lambda}(\cdot|p)$ and $\hat{\lambda}^*(\cdot|p)$, respectively, being nondecreasing on $\mathcal{I}$. (We are now working with samples of size N rather than n, and so p is of length N.) For a test at level α, use Monte Carlo methods to compute $\hat{d}$ with the property that $P\{D(\hat{p}^*) \geq \hat{d} \,|\, \mathcal{Y}\} = \alpha$. Reject H_0 if $D(\hat{p}) > \hat{d}$.

Level accuracy when the true density is $f_0(\cdot|\theta_0)$ on $\mathcal{I}$, for some θ_0, may not be sufficiently high, however, owing to the fact that the unknown parameter θ has been estimated. A standard application of the double bootstrap, to further calibrate the critical point, will often overcome this problem.

3. NUMERICAL RESULTS

3.1. Simulation study. Minimisation of power divergence distance subject to monotonicity may be accomplished discretely, by selecting ν (say) grid points and minimising the distance function $D(p)$ subject to the constraint that the derivatives at grid points are nonnegative (in the nondecreasing case) or nonpositive (in the nonincreasing case), in addition to $\sum_i p_i = 1$. We implemented this approach via the E04UCF quadratic programming routine from the NAG library.

The target density was supported on the positive half-line, and we estimated the density and hazard rate on the interval $\mathcal{I} = [0.1, 1]$, the lower endpoint being chosen to avoid edge effects. We took $\nu = 100$ equally-spaced grid points on $\mathcal{I}$, and used the Normal-reference bandwidth (Silverman, 1986, pp. 45–48) for the Epanechnikov kernel, i.e. $h_{\text{ref}} = 2.3449\, s\, n^{-1/5}$ where s denotes the sample standard deviation. We plotted the hazard function estimate $\hat{\lambda}(\cdot|\hat{p})$ at the ν grid points, but when plotting the density estimator $\hat{f}(\cdot|\hat{p})$ a finer grid was employed.

Example 1. Weibull Distribution. We simulated data from the Weibull distribution with index 2 and scale parameter 0.65, i.e. $f(x) = (2x/0.65^2)\exp\{-(x/0.65)^2)\}$, for which $\lambda(x) = 2x/0.65^2$, an increasing linear function. The parametrisation follows Johnson *et al.* (1994), and $(2, 0.65)$ were chosen because they resulted in approximately 90% of the data lying in $\mathcal{I}$. We performed 100 simulations with

$n = 50$, and took the distance function to be $D = D_\rho$ with $0 \le \rho \le 2$. When $\rho = 0$, in just 46 of the 100 simulations were the unmodified estimates of the hazard function nondecreasing on $\mathcal{I}$; that is, in just 46 out of 100 cases was it true that $\hat{p} = p_{\mathrm{unif}}$.

Figures 1 and 2 depict curves computed from those samples among the 100 that correspond to the 95th and 99th percentiles of $D_0(\hat{p})$. Estimates in the cases $\rho = 0, 1, 2$ are shown; results for other values of ρ between 0 and 2 are readily interpolated by eye. The effect of varying ρ is quite small. It can be seen that the method works particularly well in correcting the downwards trend of the unconstrained estimator at its right-hand end. Taking $\rho = 0$ discourages small weights on outlying data points; and $\rho = 2$ sometimes leads to negative probabilities unless positivity is imposed as part of the algorithm.

For the bandwidth h_{ref}, in only 1 of the 100 datasets, illustrated in Figure 3, was it possible to constrain the hazard rate in the wrong direction, i.e. to make it nonincreasing. The figure shows a substantial qualitative difference of the fitted hazard rate and the corresponding density, relative to the case where the constraint is applied in the correct direction.

The ability of our method to produce a monotone hazard rate estimate can be enhanced by reducing the bandwidth. In particular, for 55 out of the 100 simulated Weibull datasets it was possible to achieve monotonicity, with the constraint applied in the wrong direction, by reducing the bandwidth to as little as $0.20\, h_{\mathrm{ref}}$. (Reducing bandwidth further, to as low as $0.05\, h_{\mathrm{ref}}$, did not increase the number of datasets for which the constraint could be imposed.) If monotonicity was possible for bandwidths $h_1 < h_2$ then the algorithm produced monotonicity for a range of values between h_1 and h_2, but sometimes it failed to converge at some of the intermediate values. On the other hand, increasing the bandwidth above h_{ref} did not help the the algorithm to enforce the constraint in the wrong direction.

Example 2. Gamma Distribution. The gamma is another popular parametric lifetime distribution, often with a monotone increasing hazard rate on $\mathcal{I} = [0.1, 1]$. To ensure that approximately 90% of the data lay within $\mathcal{I}$ we took the gamma index to be 3 and scale to be 6. In each of 100 simulations it was possible to enforce monotonicity (in the correct direction) using our method; only 12 out of the 100 resulted in $\hat{p} = p_{\mathrm{unif}}$. The estimates corresponding to the 95th and 99th percentiles of $D_0(\hat{p})$, and to $\rho = 0$, 1 and 2, are shown in Figures 4 and 5.

For none of the 100 simulated datasets was it possible to reverse the direction of the constraint without decreasing bandwidth. However, if bandwidth was reduced to as little as $0.25\, h_{\text{ref}}$ then constraining in the wrong direction was possible in 49% of cases. (Further reduction of bandwidth did not increase this percentage.) Curiously, for one of the 100 simulated datasets it was possible to enforce the wrong constraint by increasing bandwidth (to $1.1\, h_{\text{ref}}$), but in no other instance did an increase in bandwidth enable the wrong constraint to be enforced. And in this one case, monotonicity could also be achieved by reducing bandwidth to $0.3\, h_{\text{ref}}$. Figure 6 shows our results for this dataset.

3.2. Real-data examples.

Example 3. Suicide-Study Data. For our first real-data example we used the suicide-study dataset discussed by Copas and Fryer (1980) and Silverman (1986, Chapter 2). The data represent lengths, in days, of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks. Following Silverman (1986) we worked with the logarithms of the data. We used the Epanechnikov kernel, and took the bandwidth to equal 1.1, which produces results virtually identical to those obtained using Silverman's choice of the Normal kernel and bandwidth 0.5. The resulting estimated hazard function is nondecreasing except at the left-hand end for interval $[0.5, 6]$, and this difficulty is readily removed by constraining. Figure 5 illustrates results for $\rho = 0, 1$ and 2. Curve estimates when $0 < \rho \leq 2$ are similar, although the values of $D_\rho(\hat{p})$ are larger when $\rho = 0$ ($D_0(\hat{p}) = 39.5$) than in all other cases. This occurs because small $\hat{p}_i$s incur considerable penalty when $\rho = 0$, and very small weights have to be imposed on three data values tied at 0 (after taking logarithms) in order to fulfil the monotonicity constraint. The solution when $\rho = 1$ is almost identical to that for $\rho = 2$.

Example 4. Ball Bearing Data. A dataset from Lieblein and Zelen (1956) gives the number of revolutions to failure for each of 23 ball bearings. Crowder *et al.* (1991) fitted a parametric lognormal model, i.e. the log-lifetimes were Normally distributed with moment estimates $\mu = 4.15$ and $\sigma = 0.53$ respectively. The resulting hazard function is nondecreasing over a range that includes the interval $[3, 5]$.

We estimated the hazard function on this interval using kernel methods, with the Epanechnikov kernel. The Normal-reference bandwidth, 0.67, gives a nondecreasing curve, and therefore, for illustration only, we reduced the bandwidth to see the effect of monotonisation. Decreasing h in steps of 0.01, the largest bandwidth for which monotonisation was needed was 0.47. A monotone increasing estimator was obtainable using our method until bandwidth was reduced as far as 0.29.

3.3. Hypothesis testing. Given a dataset $\mathcal{X} = \{X_1, \ldots, X_n\}$ drawn from a distribution with hazard function λ, consider testing the hypothesis H_0 that λ is nondecreasing on a given interval $\mathcal{I}$, against the complementary alternative. The approach suggested here involves estimating λ nonparametrically, subject to the constraint that it be nondecreasing, and rejecting H_0 if the observed value of $D(\hat{p})$ is too large. In our numerical work we took $D = D_0$.

Given a kernel K and bandwidth h, we calculated $\hat{p}$ and then $\widehat{F}(\cdot|\hat{p})$ as suggested in section 2.1, subject to the constraint that $\hat{\lambda}(\cdot|\hat{p})$ was nondecreasing. To determine a critical point for the test we drew bootstrap resamples $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from the distribution $\widehat{F}(\cdot|\hat{p})$, conditional on $\mathcal{X}$. For each $\mathcal{X}^*$ we computed the value $\hat{p}^*$ of p that minimised $D_0(p)$ subject to the corresponding bootstrap version of $\hat{\lambda}(\cdot|\hat{p})$ being monotone nondecreasing. The significance level of the test was taken to be the proportion of bootstrap resamples $\mathcal{X}^*$ for which $D(\hat{p}^*)$ exceeded $D(\hat{p})$.

An algorithm for generating X_j^*, an element of the bootstrap sample $\mathcal{X}^*$, involves first generating conditionally independent random integers $I(1), \ldots, I(n)$, each having the multinomial distribution that places mass $\hat{p}_i$ on i for $1 \le i \le n$. Then, conditional on $\mathcal{X}$ and the $I(j)$s, generate independent random variables $Z_1, \ldots, Z_n$ from the distribution with density K, and take $X_j^* = X_{I(j)} + h\, Z_j$.

We applied the above testing procedure to datasets simulated from exponential power distributions (Dhillon, 1981):

$$F(x) = 1 - \exp\left[1 - \exp\left\{(\alpha x)^\beta\right\}\right],$$

where $\alpha, \beta > 0$. The family of exponential power distributions can have increasing, decreasing, or so-called bathtub-shaped hazard rates whose cumulative hazard function, $-\log\{1 - F(x)\}$, is concave (or convex) on $[0, x_0]$, and convex (or concave, respectively) on $[x_0, \infty)$, for some $x_0 > 0$.

We chose $\alpha = 1.5$ and $\beta = 0.4\,(0.1)\,1.1$, a range of distributions with hazard rates increasing ($\beta = 1.0$ and 1.1), bathtub-shaped ($\beta = 0.5, \ldots, 0.9$), or decreasing ($\beta = 0.4$). The hazard functions are plotted in Figure 8(a). Sample size was $n = 50$, and 400 simulations were carried out using bootstrap simulation size 100. As in section 3.1 we used the Epanechnikov kernel and the Normal-reference bandwidth, equal to $h = 2.3449\, s\, n^{-1/5}$ in the case of a sample $\mathcal{X}$ and to $h^* = 2.3449\, s^*\, n^{-1/5}$ for a resample $\mathcal{X}^*$, where s^* denotes the standard deviation of $\mathcal{X}^*$. For those datasets with $D(\hat{p}) = 0$, no further bootstrapping is necessary since the bootstrap critical value must be positive and so the conclusion (for that dataset) must be to not reject H_0. Figure 8(b) plots the proportions of rejection at level 0.05, as a function of β. It can be seen that this empirical measure of power increases with the strength of 'non-increasingness'.

4. THEORETICAL PROPERTIES

Our first result shows that under particularly mild regularity conditions, not involving the sign of λ', there is high probability that p can be chosen so that $\hat{\lambda}'(\cdot|p)$ is always of a given sign on $\mathcal{I}$. Let $\mathcal{I} = [a, b]$ denote a compact interval, and assume f is bounded away from 0 on an open interval containing $\mathcal{I}$. Suppose too that K is a symmetric, compactly supported density with a bounded derivative on $(-\infty, \infty)$, and that $h = h(n) \to 0$ and $nh^3 \to \infty$ as $n \to \infty$.

Call these conditions (C_1). For either choice of the $\pm$ signs, taken respectively, let $\pi_\pm = \pi_\pm(n, \mathcal{I})$ denote the probability that there exists a proper probability measure $\hat{p}_\pm = (\hat{p}_{1,\pm}, \ldots, \hat{p}_{n,\pm})$ on the integers $1, \ldots, n$, with none of its marginals vanishing, such that $\pm\hat{\lambda}'(x|\hat{p}_\pm) > 0$ for all $x \in \mathcal{I}$. (The condition that no $\hat{p}_{i,\pm}$ vanish ensures that $D_0(\hat{p}_\pm)$ is finite, and hence that $D_\rho(\hat{p}_\pm)$ is well defined for each ρ. However, in formulating Theorem 4.1 we do not need to specify the distance measure D.)

Theorem 4.1. *Assume conditions* (C_1). *Then there exists $B > 0$, depending on f and on $\mathcal{I}$, such that for each choice of the $\pm$ signs,*

$$\pi_\pm = 1 - O\big\{h^{-1} \exp\big(- Bnh^3\big)\big\} \tag{4.1}$$

as $n \to \infty$.

Remark 4.1: *Size of h.* If $h = h(n) \to 0$ so slowly that

$$nh^3/\log n \to \infty \tag{4.2}$$

then the right-hand side of (4.1) converges to 1 at rate $O(n^{-B})$ for all $B > 0$. That is, the probability $\pi_\pm$ converges to 1 faster than the inverse of any polynomial in

n; we shall call this 'exponentially fast.' Of course, (4.2) is a mild assumption; in practice we would generally choose h of size $n^{-1/5}$, the optimal order for estimating a twice-differentiable f, and in that case (4.2) would be trivially satisfied. Assuming conditions (C_1) and (4.2), the following result follows from the Borel-Cantelli lemma: with probability 1, for all sufficiently large n and for either choice of the $\pm$ signs, there exists a proper probability measure $\hat{p}_\pm$, with none of its marginals vanishing, such that $\pm\hat{\lambda}'(x|\hat{p}_\pm) > 0$ for all $x \in \mathcal{I}$.

Remark 4.2: *Sign of λ'.* Note particularly that we make no assumptions about the sign of λ' on $\mathcal{I}$. In fact, we do not even require f to be differentiable, and so λ' might not be well defined. (Of course, $\hat{\lambda}'(\cdot|p)$ is well defined without λ' existing.) Alternatively, λ' could be well defined and strictly positive throughout $\mathcal{I}$, but nevertheless, as long as conditions (C_1) and (4.2) are satisfied, the probability that we can choose p such that $\hat{\lambda}'(\cdot|p)$ is strictly negative throughout $\mathcal{I}$ converges rapidly to 1 as $n \to \infty$; and of course, we can interchange 'positive' and 'negative' here.

Remark 4.3: *Uniqueness.* The theorem implies that with high probability and in general circumstances, $\hat{p}_\pm$ exists. If the distance measure D is taken to be one of the power divergence measures then uniqueness of $\hat{p}$, with probability 1 conditional on it existing, is clear from continuity of D and of the sampled distribution.

Next we discuss the convergence rate of our estimator $\hat{\lambda}(\cdot|\hat{p})$ in cases where the true hazard rate is monotone increasing. Put $\hat{p} = \hat{p}_+$, and let $\mathcal{I} = [a, b]$ be the compact interval on which we wish to estimate λ. We assume that f has two bounded, continuous derivatives on an open interval containing $\mathcal{I}$, that $f > 0$ on $\mathcal{I}$, that K is a symmetric, compactly supported density with a Hölder-continuous derivative on $(-\infty, \infty)$, that for constants $C_1, C_2 > 0$, $C_1 n^{-1/4} \le h \le C_2 n^{-1/7}$ for all sufficiently large n, and that the distance measure is D_ρ with $0 \le \rho < 1$. Call these conditions (C_2). Put $\tilde{\lambda} = \hat{\lambda}(\cdot|p_{\mathrm{unif}})$, being the naive estimator formed by taking the ratio of a standard kernel estimator of f to the corresponding estimator of $1-F$. Define $\eta_n = \{(nh^3)^{-1} \log n\}^{1/2}$, which converges to 0 under condition (C_2).

Theorem 4.2. *Assume conditions* (C_2). (*a*) *If* $\lambda' > 0$ *on* $\mathcal{I}$ *then with probability 1,* $\hat{p} = p_{\mathrm{unif}}$ *for all sufficiently large* n. *Hence, for all sufficiently large* n, $\hat{\lambda}(\cdot|\hat{p})$ *equals* $\tilde{\lambda}$. (*b*) *If* $\lambda' > 0$ *except on a finite subset* $\mathcal{S}$ *of* $\mathcal{I}$, *in an open neighbourhood of which*

λ has three continuous derivatives; and if $\lambda''' \neq 0$ on $\mathcal{S}$; then with probability 1,

$$\begin{cases} \hat\lambda(x|\hat p) - \tilde\lambda(x) = O\big(\eta_n^2 + h^4\big) & \text{if } x \notin \mathcal{S} \\ \hat\lambda(x|\hat p) - \tilde\lambda(x) = O\big(\eta_n^{7/4} h^{-1/2} + h^3\big) & \text{if } x \in \mathcal{S}\,. \end{cases} \tag{4.3}$$

Furthermore, with probability 1 the first result at (4.3) holds uniformly on any subset of $\mathcal{I}$ which excludes a certain open neighbourhood, of radius $O(\eta_n^{1/2} + h)$, of $\mathcal{S}$; and the second result holds uniformly in all $x \in \mathcal{I}$. (c) *More generally, if for some $\epsilon > 0$ the first three derivatives of λ exist and are continuous on the set*

$$\big\{x \in \mathcal{I} : \textit{for some } y \in \mathcal{I},\, \lambda'(y) = 0 \textit{ and } |x-y| \le \epsilon\big\}$$

then with probability 1, $\hat\lambda(\cdot|\hat p) - \lambda = O(\eta_n h^{-1/2} + h^{3/2})$ uniformly on $\mathcal{I}$.

A derivation of Theorem 4.2 is similar to that of Theorem 4.2 of Hall and Huang (1999), which deals with the context of monotone nonparametric regression. Therefore it is not given here.

APPENDIX: *Outline proof of Theorem 4.1*

Without loss of generality, $\mathcal{I} = [0,1]$. Choose $\delta_1 \in (0,1]$ so small that f is bounded away from 0 on $[-2\delta_1, 1+2\delta_1]$. Let $c \in (-1/100, 1/100)$ with $c \neq 0$, put $\mathcal{K} = (-\delta_1, 1+\delta_1)$, choose $|c|$ so small that $5|c| \le f(x)$ for all $x \in \mathcal{K}$, and observe that since $\delta_1 \le 1$ then $0 < 2|c| + cx \le 4|c|$ for all $x \in \mathcal{K}$. Let $\tilde p = (\tilde p_1, \ldots, \tilde p_n)$ be the random vector defined by $\tilde p_i = (2|c| + cX_i)/\{nf(X_i)\}$ if $X_i \in \mathcal{K}$, $\tilde p_i = 4/(5n)$ if $X_i \le -\delta_1$, and $\tilde p_i$ is constant (determined by the requirement that $\sum_i \tilde p_i = 1$) for all other indices i. Then $\tilde p_i \le 4/(5n)$ whenever $X_i < 1 + \delta_1$; call this result (R_1). Also, for all sufficiently large n, $E\{\hat f(x|\tilde p)\} = 2|c| + cx$ and $E\{\hat f'(x|\tilde p)\} = c$ for all $x \in \mathcal{I}$. Call this result (R_2).

Let $\mathcal{E} = \mathcal{E}(n)$ denote the event that $X_i \ge 1 + \delta_1$ for some $1 \le i \le n$. Then $P(\mathcal{E}) = 1 - O(\delta_2^n)$ for some $\delta_2 > 0$. The vector $\tilde p$ is not necessarily a distribution, but if $\mathcal{E}$ holds then it is. Assume $h < \delta_1$. Result (R_1) implies that, provided some $X_i \ge 1 + \delta_1$, the value of $1 - \widetilde F(1|\tilde p)$ equals at least 1/5. Call this result (R_3). Note that

$$\begin{aligned} \hat\lambda_1(x|p) = \{1 - \widehat F(x|\tilde p)\} \Big\{ & c + \big\{\hat f'(x|p) - E\hat f'(x|p)\big\} \\ & + \{1 - \widehat F(x|\tilde p)\}^{-1} \Big(\big[\hat f(x|p)^2 - \{E\hat f(x|p)\}^2\big] + (2|c| + cx)^2\Big)\Big\}\,. \end{aligned}$$

We may deduce from this formula, (R_2) and (R_3) that if $\mathcal{E}$ holds, and

$$\left|\hat{f}'(x|\tilde{p}) - E\hat{f}'(x|\tilde{p})\right| \leq c^2 \quad \text{and} \quad \left|\hat{f}(x|\tilde{p})^2 - \{E\hat{f}(x|\tilde{p})\}^2\right| \leq c^2 \quad \text{for all} \quad x \in \mathcal{I}\,, \tag{A.1}$$

then $\hat{\lambda}_1(x|\tilde{p}) = c\,\{1 - \widehat{F}(x|\tilde{p})\}\,\{1 + 51\,|c|\,\hat{\theta}(x)\}$, where $|\hat{\theta}(x)| \leq 1$ uniformly in $x \in \mathcal{I}$. Call this result (R_4). If $|\hat{\theta}(x)| \leq 1$ then, since $|c| < 1/100$, we have $1 + 51\,|c|\,\hat{\theta}(x) \geq 0.49$. If in addition $1 - \widehat{F}(x|\tilde{p}) > 0$ for all $x \in \mathcal{I}$ (which is the case whenever $\mathcal{E}$ holds), the sign of $\hat{\lambda}_1(\cdot|\tilde{p})$ on $\mathcal{I}$ is that of c, and of course it is possible to obtain either sign.

The theorem will follow from these results if we prove that the probability of the event defined at (A.1) equals $1 - O\{h^{-1}\exp(-Bnh^3)\}$ for some $B > 0$. This may be done by appealing to the approximation theorems of Komlós, Major and Tusn ády (1975), and using the fact that $\hat{f}'(\cdot|\tilde{p})$ and $\hat{f}'(\cdot|\tilde{p})$ are, restricted to $\mathcal{I}$, sums of independent and identically distributed random functions.

REFERENCES

BOWMAN, A.W., JONES, M.C., & GIJBELS, I. (1998). Testing monotonicity of regression. *J. Comput. Graph. Statist.*, **7**, 489–500.

COPAS, J.B. AND FRYER, M.J. (1980). Density estimation and suicide risks in psychiatric treatment. *J. Roy. Statist. Soc. Ser. A* **143**, 167–176.

CRESSIE, N.A.C. AND READ, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46**, 440–464.

CROWDER, M.J., KIMBER, A.C., SMITH, R.L. AND SWEETING, T.J. (1991). *Statistical Analysis of Reliability Data.* Chapman and Hall, London.

DHILLON, B.S. (1981). Life distributions. *IEEE Trans. Reliability.* **30**, 457-459.

GONZÁLEZ-MANTEIGA, W., CAO, R. AND MARRON, J.S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *J. Amer. Statist. Assoc.* **91**, 1130–1140.

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

HALL, P. AND HUANG, L.-S. (1999). Nonparametric kernel regression subject to monotonicity constraints. Manuscript.

HALL, P., AND PRESNELL, B. (1999). Intentionally biased bootstrap methods. *J. Roy. Statist. Soc. Ser. B* **61**, 143–158.

HARTIGAN, J.A., & HARTIGAN, P.M. (1985). The DIP test of unimodality. *Ann. Statist.*, **13**, 70–84.

HASSANI, S., SARDA, P. AND VIEU, P. (1986). Approche non paramétrique en théorie de la fiabilité: revue bibliographique. (In French.) *Rev. Statist. Appl.* **34**, 27–41.

JOHNSON, N.L., KOTZ, S. AND BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions,* Vol.1. Wiley, New York.

KAPLAN, E.L. AND MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **33**, 457–481.

KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent RV's, and the sample DF. I. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 111–131.

KOOPERBERG, C., STONE, C.J. AND TRUONG, Y.K. (1995a). Logspline estimation of a possibly mixed spectral distribution. *J. Time Ser. Anal.* **16**, 359–388.

KOOPERBERG, C. AND STONE, C.J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12**, 327–347.

KOOPERBERG, C. AND STONE, C.J. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1**, 301–328.

KOOPERBERG, C., STONE, C.J. AND TRUONG, Y.K. (1995b). Hazard regression. *J. Amer. Statist. Assoc.* **90**, 78–94.

KOOPERBERG, C., STONE, C.J. AND TRUONG, Y.K. (1995c). The L_2 rate of convergence for hazard regression. *Scand. J. Statist.* **22**, 143–157.

LIEBLEIN. J. AND ZELEN, M. (1956). Statistical investigation of the fatigue life of deep groove ball bearings. *J. Res. Nat. Bur. Stand.* **57**, 273–316.

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.

MÜLLER, H.-G. AND WANG, J.-L. (1990). Locally adaptive hazard smoothing. *Probab. Theory Related Fields* **85**, 523–538.

MÜLLER, H.-G. AND WANG, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61–76.

OWEN, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

OWEN, A.B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90–120.

PATIL, P. (1993a). Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plann. Inference* **35**, 15–30.

PATIL, P. (1993b). On the least squares cross-validation bandwidth in hazard rate estimation. *Ann. Statist.* **21**, 1792–1810.

PATIL, P., WELLS, M.T. AND MARRON, J.S. (1994). Some heuristics of kernel based estimators of ratio functions. *J. Nonparametr. Statist.* **4**, no. 2, 203–

209.

SARDA, P. AND VIEU, P. (1991). Smoothing parameter selection in hazard estimation. *Statist. Probab. Lett.* **11**, 429–434.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

SINGPURWALLA, N.D. AND WONG, M.Y. (1983). Estimation of the failure rate—a survey of nonparametric methods, Part I. Non-Bayesian methods. *Comm. Statist. A—Theory Methods* **12**, 559–588.

VILLALOBOS, M. AND WAHBA, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *J. Amer. Statist. Assoc.* **82**, 239–248.

Figure 1: *Monotonisation for data simulated from Weibull* $(2, 0.65)$ *distribution.* For the simulated dataset for which the value of $D_0(\hat{p}) = 0.90$ was at the 95th percentile, panel (a) shows the true hazard function (dashed line), the unconstrained kernel estimate with $h_{\text{ref}} = 0.28$ (dotted line), the kernel estimate under the nondecreasing constraint with $\rho = 0$ (solid line), and the versions of this quantity with $\rho = 1$ and $D_1(\hat{p}) = 0.016$ (long-dashed line) and $\rho = 2$ and $D_2(\hat{p}) = 0.58$ (extra-long dashed line). Panel (b) depicts the corresponding kernel density estimates, with the respective line types; panel (c) shows the values of $\hat{p}_i$, as a function of X_i, after the constraint has been achieved with $\rho = 0$; and panel (d) depicts the envelope-curves of values of $\hat{p}_i$, after the constraint has been achieved with $\rho = 0$, 1 and 2, again with the respective line types.

Figure 2: *Monotonisation for data simulated from Weibull* $(2, 0.65)$ *distribution.* Panel descriptions are as for Figure 1, except that now the dataset selected is the one for which $D_0(\hat{p}) = 2.3$ was at the 99th percentile, $D_1(\hat{p}) = 0.049$, $D_2(\hat{p}) = 2.3$, and $h_{\text{ref}} = 0.30$.

Figure 3: *Monotonisation in wrong direction for data simulated from Weibull* $(2, 0.65)$ *distribution.* For the one (out of 100) dataset for which monotonisation in the wrong direction was possible with $\rho = 0$ without altering the bandwidth $h_{\text{ref}} = 0.29$, panel (a) shows the true hazard function (dashed line), the unconstrained kernel estimate (dotted line), the kernel estimate under the nondecreasing constraint with $\rho = 0$ (solid line), and the version of this quantity when the constraint is changed to nonincreasing (long-dashed line). Panel (b) depicts the corresponding kernel density estimates, with the respective line types. Panel (c) shows the values of $\hat{p}_i$, as a function of X_i, after the nondecreasing constraint has been achieved with $\rho = 0$, and panel (d) shows the version of these quantities when the constraint is changed to nonincreasing. For the dataset depicted here, $D_0(\hat{p}) = 0.48$ after enforcing the nondecreasing constraint, and $D_0(\hat{p}) = 170$ after en forcing the nonincreasing constraint.

Figure 4: *Monotonisation for data simulated from Gamma* $(3, 6)$ *distribution.* Panel descriptions are as for Figure 1, except that now the dataset selected is the one for

which $D_0(\hat{p}) = 2.8$ was at the 95th percentile, $D_1(\hat{p}) = 0.06$, $D_2(\hat{p}) = 3.1$, and $h_{\text{ref}} = 0.36$.

Figure 5: *Monotonisation for data simulated from Gamma* $(3, 6)$ *distribution.* Panel descriptions are as for Figure 1, except that now the dataset selected is the one for which $D_0(\hat{p}) = 3.2$ was at the 99th percentile, $D_1(\hat{p}) = 0.03$, $D_2(\hat{p}) = 0.94$, and $h_{\text{ref}} = 0.30$.

Figure 6: *Monotonisation in wrong direction for data simulated from Gamma* $(3, 6)$ *distribution.* The dataset is the one for which monotonisation in the wrong direction was possible when either $h = 1.1\, h_{\text{ref}}$ or $0.3\, h_{\text{ref}}$; the value of h_{ref} was 0.30. Panel descriptions are as for Figure 3, except that the nonincreasing kernel estimates in panels (a) and (b) were computed with $h = 1.1\, h_{\text{ref}}$ (long-dashed line) or $h = 0.3\, h_{\text{ref}}$ (extra-long dashed line). Panel (d) shows the envelope-curves of values of $\hat{p}_i$ after the nonincreasing constraint has been achieved with the two bandwidths, again with the respective line types. For the dataset used here, $D_0(\hat{p}) = 1.1$ after enforcing the nondecreasing constraint when the bandwidth is $h = h_{\text{ref}}$, and $D_0(\hat{p}) = 129$ and 171 after enforcing the nonincreasing constraint with $h = 1.1\, h_{\text{ref}}$ and $h = 0.3\, h_{\text{ref}}$ respectively.

Figure 7: *Suicide study data set.* The dotted line, solid line, long-dashed line and extra-long dashed line in panel (a) show the unconstrained hazard rate estimate and its constrained counterparts with $\rho = 0$, $\rho = 1$ and $\rho = 2$, respectively. All curves were computed from the logged data. Panel (b) shows the corresponding density estimates, using the respective line types. Panel (c) depicts the values of weights $\hat{p}_i$ on achieving the constraint when $\rho = 0$, and panel (d) shows the envelope-curves of values of $\hat{p}_i$ after the constraint has been achieved with $\rho = 0$, 1 and 2, again with the respective line types. After achieving the constraints, $D_0(\hat{p}) = 39$, $D_1(\hat{p}) = 0.036$ and $D_2(\hat{p}) = 1.6$.

Figure 8: *Test of monotonicity for data simulated from exponential power distributions corresponding to* $(1.5, \beta)$, *for* $\beta = 0.4\,(0.1)\,1.1$. Panel (a) depicts the hazard functions, with the right-hand ends of curves corresponding to $\beta = 0.4\,(0.1)\,1.1$ in increasing order up the right-hand side of the panel. Panel (b) plots power estimates against β, for tests with nominal level 0.05 (indicated by the horizontal dotted line).