Nonparametric testing for a monotone hazard function via normalized spacings

Irène Gijbels[1] and Nancy Heckman [2]

[1] Institute of Statistics, Catholic University of Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium; gijbels@stat.ucl.ac.be

[2] Department of Statistics, The University of British Columbia, Vancouver BC, V6T 1Z2, Canada; nancy@stat.ubc.ca

November 2000

Discussion Paper 0028, Institute of Statistics
Catholic University of Louvain, Louvain-la-Neuve, Belgium
Technical Report 195,
Statistics Department, University of British Columbia

Abstract

In this paper we study the problem of testing whether a hazard function is monotonic or not. The proposed test statistics, a global test and four localized tests, are all based on normalized spacings. The global test is in fact just the test statistic of Proschan & Pyke (1967) which they introduced for testing a constant hazard function versus a non-decreasing non-constant hazard function. This global test is powerful for detecting global departures of the null hypothesis, but lacks power when there are, for example, local departures from the null hypothesis. By localizing the global test, we obtain tests which respond to this drawback. The techniques applied for localizing were inspired by recent developments in nonparametric inference. We also show how the testing procedures can be used when dealing with censored data. We evaluate the performance of the test statistics via simulation studies and illustrate them on some data sets.

Keywords and phrases: Concavity, Convexity, Decreasing failure rate, Exponential distribution, Hazard rate, Increasing failure rate, Order statistics, Random right censoring, Spacings, Type II censoring.

1 Introduction

The distribution of a variable of interest is commonly characterized via the cumulative distribution function, the probability density function or the characteristic function. When dealing with 'lifetimes' though it is often more useful to describe the distribution via the hazard function, which is the instantaneous risk of 'dying' given 'survival' up to the present time. Due to their interpretation hazard functions are a handy tool for modelling. For example, the instantaneous risk of 'dying' might be higher in one group than another. Therefore the study of lifetimes is often done via the hazard function. When talking about 'lifetime' we are automatically thinking of a positive random variable. The word 'lifetime' is used here in a broad sense: it denotes the time to the occurence of a certain specified event. Examples of events are ample: a death of an individual or the appearance of a tumor in a medical study, a failure of a machine in a factory or the 'breaking down' of a product in a quality control experiment in engineering.

Hazard functions can have various shapes. The most important ones are: monotone increasing hazards, bathtub-shaped or U-shaped hazards, monotone decreasing hazards. The human lifetime can be appropriately described by a bathtub-shaped hazard: there is an initial period with high risk due to birth defects or infant diseases, followed by a period of more or less constant risk; after a while the risk starts increasing again due to the aging process. Models with increasing hazard are frequently used. For example, manufacturers often use a 'burn-in' process for their products: the products are subjected to operation before being sold to customers. In this way, an initial early failure of defective items is prevented and the items which are put on the market exhibit gradual aging. Decreasing hazards can be useful when modelling survival time after a successful medical treatment. More examples of typical hazard functions are provided in, for example, Lawless (1982). See also this reference for inference procedures for the most important parametric models.

Early references on nonparametric estimation of a hazard function are Watson & Leadbetter (1964a,b) who introduced kernel estimators for the hazard function, and Nelson (1969) who focused on estimating the cumulative hazard function. The literature on nonparametric estimation of a hazard function under monotonicity constraints is rather limited. Maximum likelihood estimators for an increasing hazard function were studied by Grenander (1956), Marshall & Proschan (1965), Padgett & Wei (1980), Mykytyn &

Santner (1981), Wang (1986), Tsai (1988), Mukerjee & Wang (1993) and Huang & Wellner (1995). Recently, Hall et al. (2000) proposed a method for monotonising kernel type estimators of a hazard function. The method is based on a 'biased bootstrap'; see Hall & Presnell (1999).

In this paper we study the problem of testing whether a hazard function is monotonic or not. If we reject the hypothesis of monotonicity, then we should not consider modelling survival times via a certain parametric family, such as the Weibull, which always has a monotone hazard. On the other hand, if we can accept the hypothesis that the hazard is monotone, then we may more accurately estimate it using the appropriate monotonocity constraint or by using an appropriate parametric family of distributions. Furthermore, confirming or disproving monotonocity is of interest in its own right: if we believe that the hazard function of time to relapse after a treatment is decreasing, then we know that the positive effect of the treatment increases over the time span considered.

Most papers in the literature focus on testing the null hypothesis of a constant hazard rate versus the alternative of a non-decreasing, non-constant, hazard rate:

$$\begin{array}{rll} & H_0^{\mathrm{R}}: & \text{the hazard function is constant} \\ \text{versus} & H_1^{\mathrm{R}}: & \text{the hazard function is a non-decreasing, non-constant, function.} \end{array} \tag{1.1}$$

This is a rather restricted testing problem. The null hypothesis is equivalent to testing that lifetimes are exponentially distributed and the alternative hypothesis only involves a rather small class of non-constant functions. In this paper we generalize both the null and alternative hypotheses. Specifically, we consider testing

$$\begin{array}{rll} & H_0: & \text{the hazard function is non-increasing} \\ \text{versus} & H_1: & \text{the hazard function increases on some interval,} \end{array} \tag{1.2}$$

but the methods developed in this paper can also be applied for testing

$$\begin{array}{rll} & H_0: & \text{the hazard function is non-decreasing} \\ \text{versus} & H_1: & \text{the hazard function decreases on some interval.} \end{array} \tag{1.3}$$

In many situations observations are censored, that is we do not observe the particular event defining the lifetime. For example, an individual might drop out of a medical study before the actual end of the study and hence one only knows that his lifetime is larger than the observed time. Censoring complicates but does not preclude inference on the distribution of the survival times.

In the uncensored case seminal work on testing for a constant hazard rate versus a non-decreasing, non-constant, hazard rate has been done by Proschan & Pyke (1967), Bickel & Doksum (1969) and Bickel (1969). These tests are based on normalized spacings, which we define in equation (2.3). Barlow & Proschan (1969) then extended these tests to handle Type I censored data. Other tests for the complete data case were proposed by Barlow & Doksum (1972), Ahmad (1975) and Klefsjö (1983), among others. For randomly censored data Gerlach (1987) introduced a test statistic based on a measure of the deviation of the lifetime distribution towards a lifetime distribution with increasing failure rate. Kumazawa (1992) generalized this, and proposed a general class of statistics by introducing weight functions. Related references on goodness-of-fit testing for exponentiality are Gail and Gastwirth (1978a,b), among others. More recently Hall et al. (2000) proposed a nonparametric test based on evaluating the distance between the monotonised estimator and the standard kernel estimator. This test depends on the smoothing parameter involved in the kernel estimator.

Our testing procedures are localized versions of Proschan & Pyke's (1967) test of (1.1), who based their test on comparisons of all normalized spacings. We show in Section 3 that their test can also be used for testing the null hypotheses in (1.2) and (1.3). While their test is powerful for the alternative hypothesis of a monotone hazard, it is not designed to test against the more general alternative hypotheses of (1.2) and (1.3). Our local versions of their test do not focus on an overall comparison of spacings, but rather are based on comparisons belonging to certain 'local' regions. Our ideas are inspired by recent developments in nonparametric testing using local techniques such as those of Gijbels et al. (2000) and Hall & Heckman (2000). Another approach to localize tests can be found in Dümbgen (1999). The advantage of our tests in comparison to the recently proposed nonparametric test of Hall et al. (2000) is that our tests do not depend on a smoothing parameter.

Based on simulation studies, we find that our local tests of (1.2) are more powerful than the global test when the true hazard is decreasing except on a small interval, where it is increasing. When the hazard is increasing except for a small interval of decrease the local methods are comparable to or slightly better than the global test. When the hazard is strictly increasing, the global test is most powerful, as expected.

The paper is organized as follows. In Section 2 we briefly describe some preliminary facts about hazard functions which will be necessary for a clear explanation of the testing procedures. These procedures are introduced in Section 3, and evaluated via a simulation study in Section 4. In the same section we also illustrate our procedures on a real data set. An approach to extend the testing procedures to the censored data case is provided in Section 5, along with a small simulation study and analysis of a data set.

2 Decreasing and increasing hazard functions

Let T be the nonnegative variable of interest, and denote by F the distribution function of T. Assume for a moment that T is a continuous random variable with density f. The hazard function of T is then defined for values of t with $F(t) < 1$ as

$$h(t) = \lim_{\Delta \to 0+} \frac{\text{pr}\{t \leq T < t + \Delta | T \geq t\}}{\Delta}.$$

Thus $h(t)$ represents the risk that an individual or item fails immediately after time t given survival up to time t. The cumulative hazard function is defined via the survival function $S(t) = 1 - F(t) = \text{pr}\{T > t\}$:

$$H(t) = -\log S(t),$$

but can also be defined as $H(t) = \int_0^t h(u)du$. The hazard function then equals $h(t) = -d\{\log S(t)\}/dt = f(t)/\{1 - F(t)\}$. Similar definitions of the hazard function and the cumulative hazard function can be given in the case of a discrete random variable T or in case of a mixed discrete-continuous random variable. Our method is valid in general, for T any nonnegative random variable.

We first recall a simple fact about the hazard and the cumulative hazard function which will allow us to formulate testing problems which are equivalent to problems (1.2) and (1.3) but which do not require T to have a density. Proposition 1 states that the monotonicity of the hazard function $h(\cdot)$ can be formulated in terms of the convexity/concavity of the cumulative hazard function $H(\cdot)$. A proof of this proposition can be found in Barlow & Proschan (1965).

Proposition 1.
The following statements are equivalent:

(i). $h(t)$ is a non-decreasing (respectively non-increasing) function

(ii). $H(t)$ is a convex (respectively concave) function on $\{t \geq 0 : F(t) < 1\} \equiv [0, \tau_F[$,

where τ_F denotes the upper endpoint of the support of F, i.e. $\tau_F = \sup_t \{t \geq 0 : F(t) < 1\}$.

As a consequence of this proposition testing problem (1.2) is equivalent to

$$\begin{aligned} & H_0 : \ H \text{ is a concave function on } [0, \tau_F[\\ \text{versus} \quad & H_1 : \ H \text{ is not a concave function,} \end{aligned} \tag{2.1}$$

and testing problem (1.3) is equivalent to

$$\begin{aligned} & H_0: \ H \text{ is a convex function on } [0, \tau_F[\\ \text{versus} \quad & H_1: \ H \text{ is not a convex function.} \end{aligned} \tag{2.2}$$

We will focus on testing problem (1.2) or (2.1) since this will allow us to stay close to the original description of the test of Proschan & Pyke (1967). In Remark 2 in Section 3 we will indicate how to use our test procedures for testing (1.3) or equivalently (2.2).

We now briefly describe the test of Proschan & Pyke (1967), proposed for testing (1.1), and justify later on that it can be used for testing the null hypothesis in (1.2). Note first that the null hypothesis H_0^{R} is equivalent to H_0^{R}: T has an exponential distribution. See for example Lawless (1982). Furthermore, the null hypothesis in (1.2) is a composite hypothesis, including the case of the exponential distribution.

Let $T_1, \ldots, T_n$ be a sample of independent and identically distributed random variables from T, and denote by $0 \equiv T_{(0)} \leq T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$ the order statistics of the observed times. The test of Proschan & Pyke (1967) is based on the normalized spacings

$$\bar{D}_{ni} = (n - i + 1)(T_{(i)} - T_{(i-1)}), \qquad i = 1, \ldots, n, \tag{2.3}$$

and uses the test statistic

$$V_n = \sum_{\substack{i,j=1 \\ i<j}}^{n} V_{ij} \qquad \text{with} \quad V_{ij} = I\{\bar{D}_{ni} > \bar{D}_{nj}\}. \tag{2.4}$$

The null hypothesis is rejected when the observed value of V_n is too large. Our tests will also be based on the normalized spacings, but will not compare all of them in a global sense as is done in (2.4). We will be more selective in comparing the normalized spacings and work more locally.

To understand this test statistic one needs to have insights into the behaviour of the normalized spacings. Those have been studied in detail in the literature. See for example Barlow & Proschan (1965), Appendix 2, and earlier references provided therein, as well as Robertson et al. (1988). Some known facts about the normalized spacings are summarized in Proposition 2 below. Recall that a random variable U is said to be stochastically smaller (larger) than a random variable W if $\mathrm{pr}\{U > u\} \leq \mathrm{pr}\{W > u\}$ for all u (respectively $\mathrm{pr}\{U > u\} \geq \mathrm{pr}\{W > u\}$ for all u). We use the notation $U \leq_{st} W$ ($U \geq_{st} W$) for indicating that U is stochastically smaller (larger) than W. The proof of item (i) below can be found in Epstein & Sobel (1953), whereas an easily accessible proof of items (ii) and (iii) is given in, for example, Robertson et al. (1988).

Proposition 2.

(i). If the T_i's are exponentially distributed with mean θ then $\bar{D}_{n1}, \ldots, \bar{D}_{nn}$ are independent exponential random variables with mean θ.

(ii). If the T_i's have a non-increasing hazard function then $\bar{D}_{n1} \leq_{st} \bar{D}_{n2} \leq_{st} \cdots \leq_{st} \bar{D}_{nn}$.

(iii). If the T_i's have a non-decreasing hazard function then $\bar{D}_{n1} \geq_{st} \bar{D}_{n2} \geq_{st} \cdots \geq_{st} \bar{D}_{nn}$.

An immediate consequence of Proposition 2 (*i*) is that, for $i, j = 1, \ldots, n$, $i \neq j$, the V_{ij}'s are Bernoulli distributed with $\text{pr}\{V_{ij} = 1\} = 1/2$ if the T_i's are exponentially distributed. ¿From Proposition 2 (*ii*) and (*iii*) we might suspect that, for $i < j$, $\text{pr}\{V_{ij} = 1\}$ is less than $1/2$ if the T_i's have a decreasing hazard function and is greater than $1/2$ if the T_i's have an increasing hazard function. This is indeed the case, as shown by Proschan & Pyke (1967) and implied by Theorem (*ii*) and (*iii*) in the following section. Motivated by these facts Proschan & Pyke (1967) proposed rejecting H_0^{R} in favour of H_1^{R} when V_n is too big. Determining the critical region for this test is straightforward in principle since the distribution of the V_{ij}'s is determined under the null hypothesis H_0^{R} of exponentiality.

3 Test statistics

We now consider several testing procedures for testing problem (1.2). All of our test statistics are non-decreasing functions of the V_{ij}'s, $i < j$. We prove below that we can calculate the significance levels of our tests using rejection probabilities calculated for exponentially distributed lifetimes. We also argue heuristically that our tests will have some power against the alternative hypothesis. To accomplish this we first transform the lifetimes to exponential random variables and then compare the V_{ij}'s from the original lifetimes to the V_{ij}'s of the transformed lifetimes. Specifically let

$$T_i^0 = H(T_i) = -\log S(T_i), \qquad i = 1, \ldots, n, \tag{3.1}$$

and let V_{ij}^0 be the indicator variable, defined similarly as V_{ij} but based on the T_i^0's. The proof of the following theorem is similar to that of Proschan & Pyke (1967), but is given here briefly, for clarity.

Theorem.

(i). The random variables $T_1^0, \ldots, T_n^0$ are independent and identically distributed having an exponential distribution with unit mean.

(ii). If H is concave on an interval $[a, b[$, then $V_{ij}^0 \geq V_{ij}$ for $i < j$, $i, j = 1, \ldots, n$, if $T_{(i-1)}$ and $T_{(j)}$ are in the interval $[a, b[$.

(iii). If H is convex on an interval $[a, b[$, then $V_{ij}^0 \leq V_{ij}$ for $i < j$, $i, j = 1, \ldots, n$, if $T_{(i-1)}$ and $T_{(j)}$ are in the interval $[a, b[$.

Remark 1. When testing whether the hazard function is decreasing on the whole domain, we consider $[a, b[$ equal to $[0, \tau_F[$. See Proposition 1. If, on the other hand, we want to test whether the hazard function is decreasing on a certain subset of its domain then we will choose the interval $[a, b[$ accordingly.

Proof of the Theorem.

Proof of statement (i).

The random variables $T_1^0, \ldots, T_n^0$ are independent since $T_1, \ldots, T_n$ are independent. Moreover, it is easy to show that T_i is exponentially distributed with unit mean. See also Proschan & Pyke (1967) and Barlow & Proschan (1965, 1975).

Proof of statement (ii).

If H is concave on $[a, b[$, then this implies that for all $i < j$

$$\frac{H(T_{(i)}) - H(T_{(i-1)})}{T_{(i)} - T_{(i-1)}} \geq \frac{H(T_{(j)}) - H(T_{(j-1)})}{T_{(j)} - T_{(j-1)}}$$

whenever $T_{(i-1)}$ and $T_{(j)}$ are in $[a, b[$. Since H is a non-decreasing function, the ordering of the T_i^0's is the same as that of the T_i's. Let $T_{(1)}^0 \leq T_{(2)}^0 \leq \cdots \leq T_{(n)}^0$ be the ordered T_i^0's. Thus

$$\frac{T_{(i)}^0 - T_{(i-1)}^0}{T_{(j)}^0 - T_{(j-1)}^0} \geq \frac{T_{(i)} - T_{(i-1)}}{T_{(j)} - T_{(j-1)}},$$

or also

$$\frac{\bar{D}_{ni}^0}{\bar{D}_{nj}^0} \geq \frac{\bar{D}_{ni}}{\bar{D}_{nj}}.$$

Consequently, if $V_{ij} = 1$, then $V_{ij}^0 = 1$. If $V_{ij} = 0$, then V_{ij}^0 can equal 0 or 1. Hence, in general, $V_{ij}^0 \geq V_{ij}$, for all $i < j$.

Proof of statement (iii).
This is similar to the proof of statement (ii) but relies on the convexity of H. □

The following Corollary follows immediately.

Corollary.

Let $T_1, \ldots, T_n$ be independent and identically distributed survival times, T_i^0 as defined in (3.1), let V be a non-decreasing function of the V_{ij}'s, $i < j$, with V_{ij} as defined in (2.3) and (2.4), and let V^0 be the corresponding statistic based on the T_i^0's.

(i). If H is concave on an interval $[a, b[$, and all the T_i's fall in this interval, then $V^0 \geq V$.

(ii). If H is convex on an interval $[a, b[$, and all the T_i's fall in this interval, then $V^0 \leq V$.

Suppose that, in the testing problem (1.2), we reject H_0 when $V > c$, where V satisfies the conditions of the Corollary. We now easily see that the significance level α can be calculated assuming lifetimes have an exponential distribution. Indeed, for all non-increasing hazard functions h or, equivalently, all concave cumulative hazard functions H, we find from Corollary (i) that $\text{pr}\{V > c\} \leq \text{pr}\{V^0 > c\}$, for all c, and since a constant hazard is just an element of this class of non-increasing hazard functions, we have

$$\alpha = \sup_{h \text{ non-increasing}} \text{pr}\{V > c\} = \text{pr}\{V^0 > c\}.$$

Thus, to calculate the critical region of the test or the p-value we simulate samples of independent and identically distributed random variables from an exponential distribution with unit mean, and calculate the statistic V^0.

Corollary (ii) gives us a heuristic indication that our test might have power if the hazard is increasing on some interval $[a, b[$: we would expect our test statistic V to be large because of that interval of increase.

Remark 2. It is easy to see that our generic test statistic V can also be used for testing problem (1.3). Indeed, from part (iii) of the Theorem, we see that, for H convex and $i > j$, $V_{ij}^0 \geq V_{ij}$. So it suffices to work with the original test statistic V but with the reverse of the normalized spacings, $\bar{D}_{ni}^I \equiv \bar{D}_{n,n-i+1}$.

We consider five tests, all of which are based on test statistics which are non-decreasing functions of the V_{ij}'s for $i < j$, and we reject the null hypothesis when the test statistic is large. Our first statistic is V_n, as proposed by Proschan & Pyke (1967), which we call the *global sign test*, denoted by

$$\mathcal{T}_{\text{global}} \equiv V_n = \sum_{\substack{i,j=1 \\ i<j}}^{n} V_{ij}. \tag{3.2}$$

The other four test statistics are localized versions of $\mathcal{T}_{\text{global}}$.

A first local version is obtained by summing not all indicator variables, but using a local sum and taking the maximum over these local sums. More precisely, for two indices s and k, with values in the set $\{1, \ldots, n\}$ and such that $s + k \leq n$, define the local sum

$$\mathcal{V}_{sk} = \sum_{s \leq i < j \leq s+k}^{n} V_{ij}.$$

Note that $\mathcal{T}_{\mathrm{global}} = \mathcal{V}_{1,n-1}$. Before taking a maximum over s and k, we normalize $\mathcal{V}_{sk}$ to take account of the number of terms in the sum. Proschan & Pyke (1967) provided the mean and variance of the statistic $\mathcal{T}_{\mathrm{global}}$ under H_0^{R} (i.e. exponentiality of the T_i's): $E(\mathcal{T}_{\mathrm{global}}) = n(n-1)/4$ and $\mathrm{var}(\mathcal{T}_{\mathrm{global}}) = (2n+5)(n-1)n/72$. This result was obtained by Mann (1945). The mean and variance of the quantity $\mathcal{V}_{sk}$ are easily obtained from these expressions by taking $n = k+1$. So, if the T_i's are exponentially distributed we have

$$\mu_k \equiv E(\mathcal{V}_{sk}) = \frac{(k+1)k}{4} \qquad\qquad v_k \equiv \mathrm{var}(\mathcal{V}_{sk}) = \frac{(2k+7)(k+1)k}{72}.$$

Let $\mathcal{V}^*_{sk}$ denote the normalized local sums: $\mathcal{V}^*_{sk} = v_k^{-1/2}\,(\mathcal{V}_{sk} - \mu_k)$. Then the *local sign test* is defined as

$$\mathcal{T}_{\mathrm{local}} = \max_{1 \leq s \leq n-1} \; \max_{1 \leq k \leq n-s} \mathcal{V}^*_{sk}. \tag{3.3}$$

Our other three tests use comparisons of spacings that are separated by a certain index, say k with $k \geq 1$. More precisely, define

$$\mathcal{V}_{sk}(r) = \sum_{i=0}^{r} V_{s+i,s+i+k}\ .$$

The three tests differ in the way we normalize the $\mathcal{V}_{sk}(r)$'s and in the values of s, k and r used when taking the maximum.

For one test, we use the fact that, if the T_i's are independent and exponentially distributed random variables, and hence the normalized spacings $\bar{D}_{ni}$'s are independent and identically distributed exponential random variables, then

$$\mu(r) \equiv E\{\mathcal{V}_{sk}(r)\} = \frac{r+1}{4} \qquad\qquad v_k(r) \equiv \mathrm{var}\{\mathcal{V}_{sk}(r)\} = \frac{r+1}{4} - \frac{(r+1-k)_+}{6},$$

where $y_+ = \max(y, 0)$. The calculation of $\mu(r)$ is straightforward since $\mathcal{V}_{sk}(r)$ is a sum of (possibly) dependent Bernoulli random variables with success probabilities $1/2$. The calculation of $v_k(r)$ involves the calculation of covariances among the V_{ij}'s and careful counting of similar terms. Let $\mathcal{V}^*_{sk}(r)$ denote the normalized quantity: $\mathcal{V}^*_{sk}(r) = v_k(r)^{-1/2}\,\{\mathcal{V}_{sk}(r) - \mu(r)\}$. We then define the test statistic

$$\mathcal{T}_{\mathrm{skip}} = \max_{1 \leq s \leq n-1} \; \max_{1 \leq k \leq n-s} \; \max_{0 \leq r \leq n-k-s} \mathcal{V}^*_{sk}(r), \tag{3.4}$$

which we call the *local skip test* since it skips spacings.

Another statistic is obtained by approximating the variance of $\mathcal{V}_{sk}(r)$ assuming that all terms in the sum are independent Bernoulli random variables with success probability 1/2. We thus obtain the approximated variance

$$\tilde{v}(r) = \frac{r+1}{4}.$$

The local skip test using this approximated variance will be based on $\tilde{\mathcal{V}}^*_{sk}(r) = \tilde{v}(r)^{-1/2}$ $\{\mathcal{V}_{sk}(r) - \mu(r)\}$, and leads to the test statistic

$$\mathcal{T}_{\text{skipapprox}} = \max_{1\leq s\leq n-1} \max_{1\leq k\leq n-s} \max_{0\leq r\leq n-k-s} \tilde{\mathcal{V}}^*_{sk}(r), \tag{3.5}$$

which we will call the *approximate local skip test.*

Our last test statistic only considers $\mathcal{V}_{sk}(r)$ with $r \leq k+1$. Note that, in this case and assuming exponentially distributed lifetimes, $\mathcal{V}_{sk}(r)$ is indeed a sum of independent Bernoulli distributed random variables. Also, $\tilde{v}(r) = v_k(r)$. Our *restricted local skip test* is defined as

$$\mathcal{T}_{\text{skiprest}} = \max_{1\leq s\leq n-1} \max_{1\leq k\leq n-s} \max_{0\leq r\leq \min(k+1,n-k-s)} \tilde{\mathcal{V}}^*_{sk}(r). \tag{3.6}$$

4 Analysis with uncensored data

4.1 Simulation study

In this section we evaluate via a simulation study the performance of the five tests discussed in Section 3 for testing problem (1.2).

In our simulations we model the logarithm of the hazard function as

$$\log h(t) = a_0 + a_1 \log t + \beta(2\pi\sigma^2)^{-1/2} \exp\{-\frac{(t-\mu)^2}{2\sigma^2}\}. \tag{4.1}$$

The first two terms are the logarithm of the Weibull hazard, and are obviously increasing (decreasing) when a_1 is positive (negative). The third term induces a bump at location $t = \mu$ and thus when $\beta \neq 0$, the hazard is not monotone. When $a_1 = \beta = 0$, (4.1) is the exponential log hazard. Figures 1 and 2 illustrate properties of the model for several values of parameters: $a_0 = 0$, $\mu = 1$, $a_1 = -0.2$ or 0.1, $\beta = 0$ or 0.3, $\sigma = 0.1$ or 0.2. Figure 1 shows the densities, the hazards and the cumulative hazards, the first and last calculated numerically. Figure 2 shows plots of the 5th, 25th, 50th, 75th and 95th percentiles of the normalized spacings $\bar{D}_{ni}$ as a function of i, for the case that $n = 50$. These percentiles were calculated via 1000 simulated data sets. In general, when a_1 is positive the hazard has an overall increasing trend and the normalized spacings have a decreasing trend. When a_1 is negative the opposite holds. The trend is more pronounced for large values of $|a_1|$. Decreasing σ or increasing β produces a sharp peak in the hazard function or equivalently a dip in the normalized spacings.

FIGURES 1 AND 2 APPEAR HERE.

For each parameter configuration considered, we generated 1000 data sets of size 50 and calculated our five test statistics. Recall from Section 3 that our p-values can be calculated assuming that the 50 lifetimes are exponentially distributed. We calculated our p-values using 2000 simulated exponential data sets. Table 1 contains the proportion of times out of 1000 that H_0 is rejected. The standard errors of these proportions are approximately 0.01. The only configurations of parameters that are contained within the null hypothesis are $\beta = 0$ with $a_1 = -0.2$, -0.1 or 0.0. Of the local statistics, the restricted local skip test in (3.6) and the approximate local skip test in (3.5) are almost never less powerful than the local skip test in (3.4) and the local sign test in (3.3), and are usually significantly more powerful. Note that the performances of the restricted local skip test and the approximate local skip test are almost identical. Indeed, closer inspection of the simulation study shows that these two statistics agree about 80% of the time, because the maximum value of the normalized $\mathcal{V}_{sk}(r)$'s occurs when $r \leq k+1$. The global sign test in (3.2) is powerful when the overall trend of the hazard is increasing ($a_1 > 0$), but its performance relative to the local tests deteriorates as we move further from the Weibull. For values of $a_1 \leq 0$ and $\beta \neq 0$, we see that the global sign test is less powerful than the local tests. Thus the local tests are better than the global sign test at detecting the bump in a decreasing log hazard. The differences between the global test and the local tests increase as the value of $|a_1|$ increases.

TABLE 1 APPEARS HERE.

Our simulation results actually hold for a family of parameters, by the following argument. The V_{ij}'s, and hence our statistics, are scale invariant, that is they are not changed if the lifetimes are multiplied by a positive constant. Thus our simulations for lifetimes with survival function $S(\cdot)$ hold for data with survival function $S^*(\cdot) = S(\cdot/\gamma)$, $\gamma > 0$. Translating this condition into conditions on the hazard function, we see that our results for parameter values $a_0, a_1, \beta, \sigma, \mu$ also hold for parameter values $a_0^* = a_0 - (a_1 + 1)\log\gamma$, $a_1^* = a_1$, $\beta^* = \gamma\beta$, $\sigma^* = \gamma\sigma$ and $\mu^* = \gamma\mu$.

We also conducted a simulation study to compare our testing methods to the simulation results of Hall et al (2000). These authors consider testing problem (1.3) and they simulate data from a model with hazard $h(t) = \alpha\beta(\alpha t)^{\beta-1}\exp\{-(\alpha t)^{\beta}\}$ with $\alpha = 1.5$ and

$\beta = 0.4(0.1)1.1$. We consider sample size $n = 50$ and tests of level 0.05. When $\beta = 0.4$, the hazard function is decreasing, so we should easily reject H_0. Hall et al reject H_0 91.75% of the time, and our local methods are comparable with, for example, the restricted local skip method rejecting 94.5% of the time. As expected for decreasing hazards, the global method is very powerful, rejecting 99.5% of the time. When $\beta = 1.0$ or 1.1, the hazard function is increasing and therefore we should not reject the null hypothesis. Here, our tests outperform Hall et al's. The global test never rejects H_0, the restricted local skip test and the approximate local skip test reject less than 1% of the time, and our other two tests reject 1 or 2% of the time. In contrast Hall et al's test rejects 6.5% or 5.25% of the time. When $\beta = 0.5, ..., 0.9$ the hazard function is bathtub shaped. All methods lose power as β increases, but our methods lose power more rapidly than Hall et al's. In fact, when $\beta = 0.9$, our methods perform poorly, rejecting less than 3% of the time, compared to Hall et al's rejection rate of 8%. However, when $\beta = 0.9$ the bathtub shaped hazard is mainly decreasing and the normalized spacings are mainly increasing. When we try our methods to test the null hypothesis that the hazard function is non-decreasing we find that the global test rejects the null hypothesis 55.5% of the time and the restricted local skip test rejects 31% of the time.

4.2 Illustration with real data

We now study the hazard function for the lifetime of ball bearings. The data set, originally discussed in Lieblein & Zelen (1956), can be found in Lawless (1982), page 228, and consists of the number of revolutions (in millions) before failure of 23 ball bearings. Leiblein & Zelen analyzed the data with a Weibull model, and Lawless considered the Weibull and the log normal model, which has a hazard that increases then decreases.

Figure 3 contains a plot of the cumulative hazard, calculated as the negative of the logarithm of 1 minus the empirical distribution function. The plot on the right side of the figure gives the normalized spacings $\bar{D}_{ni}$ versus i. ¿From Figure 3, we see that the cumulative hazard might be convex and that the normalized spacings have an overall decreasing trend. We first test (1.2) and then test (1.3) as described in Remark 2 of Section 3. The p-values are presented in Table 2. Both the global and local sign tests reject the hypothesis that the cumulative hazard is concave. All tests give very high p-values for testing the null hypothesis that the cumulative hazard is convex. We conclude that the cumulative hazard function is not concave but may be convex.

FIGURE 3 AND TABLE 2 APPEAR HERE.

5 Censored case

In practice, often the T_i's are not completely observed, i.e. some of them are censored. When data are Type II censored, our methods easily apply. Type II censoring occurs when a study continues until r out of the n items under study have failed. Hence at the end of the study the surviving $n-r$ items are censored and we only observe the first r order statistics of the lifetimes, $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(r)}$. We can still form the first r normalized spacings and can calculate a statistic based on these. We are thus using a test statistic that is a non-decreasing function of the V_{ij}'s, $1 \leq i < j \leq r$, and hence is also a non-decreasing function of the V_{ij}'s, $1 \leq i < j \leq n$. Therefore the results of Section 3 still hold. We can calculate our p-values using the normalized spacings from the first r order statistics from an exponential distribution or, equivalently, by Proposition 2 (i), from a sample of r exponentials.

Of course, in a parametric analysis in a Type II censoring problem, we would not ignore the fact that there are $n-r$ lifetimes greater than $T_{(r)}$. However, in a nonparametric setting, this fact gives us no information. We can only make inferences about the hazard on the range of observed data. We can use our methods to test the null hypothesis that h is decreasing on an interval $[a, b[$, where b is a fixed point such that $T_{(r)} > b$ with probability tending to one.

In the case of random right censoring each of the survival times T_i is subject to a censoring time C_i, which is usually assumed to be independent of T_i. The observations consist of the pairs (X_i, δ_i), $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$, the observed time, and $\delta_i = I\{T_i \leq C_i\}$. In order to use our test statistics based on normalized spacings, we propose imputing lifetimes for the censored values. We then calculate our test statistics on the "complete" data set and calculate p-values as before, using the exponential distribution and ignoring the effect of imputation. We impute the lifetimes based on a Weibull distribution with survival function $S(t) = \exp\{-(\lambda t)^{a_1+1}\}$, with parameters estimated from the original censored data. Note that this model is just a reparametrization of (4.1) with $\beta = 0$. We compute the maximum likelihood estimates of the Weibull parameters and denote them by $\hat{\lambda}$ and $\hat{a}_1$. See Lawless (1982) for parametric inference from randomly right censored data. If X_i is a censored time, then we replace it with X_i^*, randomly drawn from the conditional distribution of a Weibull with parameters $\hat{\lambda}$ and $\hat{a}_1$, given that the Weibull is greater than X_i. The Weibull survival function is $\exp\{-(\hat{\lambda}t)^{\hat{a}_1+1}\}$ and thus, given $\hat{\lambda}$, $\hat{a}_1$ and X_i, X_i^* has survival function $S^*(t) = \exp\{-(\hat{\lambda}t)^{\hat{a}_1+1} + (\hat{\lambda}X_i)^{\hat{a}_1+1}\}$.

Our reasoning in the above is that, for uncensored lifetimes, our significance levels can be calculated assuming the lifetimes are exponential. Thus for censored data, to try to maintain the appropriate significance level, one might consider imputing the censored observations via an exponential distribution and then calculating the p-value assuming the sample was complete. However, generating from an exponential distribution, which has a constant hazard, might lead to a test with little power. We should use a family which includes the exponential distribution but is more flexible, including both increasing and decreasing hazards. Noting that an exponential distribution is just a Weibull distribution with $a_1 = 0$, we hope to improve our power but still maintain our significance level, by allowing for estimation of the parameter a_1.

We conducted a small simulation study to evaluate the performance of this approach, but just comparing the global sign test and the best local test, the restricted local skip test in (3.6). We used the log hazard as in (4.1) and the same values of the parameters as in the simulation study of Section 4. Our censoring variables were independent of the lifetimes and were exponentially distributed with mean chosen to yield a 20% censoring rate. As in Section 4, our sample size was 50 and we generated 1000 data sets for each parameter configuration. The results are displayed in Table 2.

TABLE 2 APPEARS HERE.

If our conjecture concerning our procedure were correct, we would expect that, when lifetimes are exponentially distributed, a 5% level test would reject 5% of the time. From the table entry for $a_1 = \beta = 0$, we see that this is only approximately true.

The main difference between these simulations results and the ones for uncensored data is that the global test does well. This is most likely because we are imputing using data that have a monotone hazard and the global test does very well at distinguishing between a decreasing hazard and an increasing hazard. For instance, when $a_1 = 0.1$ our estimates of the Weibull parameters tend to be positive. When there is no censoring, the interspacings have a downward trend. See Figure 1. With censoring, this downward trend is enhanced by imputation from a Weibull with increasing hazard and thus the global test gains power.

We apply our tests to the electrical appliance data set, which can be found in Lawless (1982), page 7, but originally appeared in Nelson (1970). Thirty-six appliances were operated repeatedly by an automatic testing machine, with lifetimes recorded as the number of cycles of use completed until failure. Three of the appliances were still operating at the end of the study. Since failure time is measured in terms of number of cycles, we treat these three appliances as being randomly right-censored. We impute these three missing values as described above. Figure 4 contains a plot of the estimated cumulative hazard, equal to minus the logarithm of the Kaplan-Meier estimate of the survival function, based on the original data set. On the right side of Figure 4 is a plot of the normalized spacings

$\bar{D}_{ni}$ versus i, based on the data set containing the three imputed values. The diamonds indicate normalized spacings affected by the imputation: the three imputed values correspond to order statistics $T_{(21)}, T_{(33)}$ and $T_{(36)}$, and thus are used to compute $\bar{D}_{ni}$ for $i = 21, 22, 33, 34$ and 36.

From the left plot in Figure 4, we see that the cumulative hazard appears as neither convex nor concave. This is supported by the fact that the normalized spacings have no monotone trend. Our tests partially confirm this. We first test (1.2) and then test (1.3) as described in Remark 2 of Section 3. The p-values are presented in Table 4. From our local tests, clearly we can reject the hypothesis that the hazard is non-increasing. We see that the global test has little power here. We conclude that the hazard function is not decreasing but perhaps is increasing.

The largest observed time may be influential in the analysis, being much larger than the next largest observed time. In addition, the largest observed value is a censoring time. Therefore we redid the entire analysis omitting this influential point. The results are similar except that the global test has even less power and the null hypothesis of (1.3) is less plausible, with p-values for this test equal to 0.780 (global sign test), 0.449 (restricted local skip test), 0.296 (approximate local skip test), 0.056 (local skip test), and 0.1525 (local sign test).

FIGURE 4 AND TABLE 3 APPEAR HERE.

Acknowledgements

This research was supported by 'Projet d'Actions de Recherche Concertées', No. 98/03–217 of the Belgian government and by the Natural Sciences and Engineering Research Council of Canada. The second author would like to thank the Institute of Pure and Applied Mathematics and the Institute of Statistics of the Université Catholique de Louvain, for the financial support and their hospitality.

References

AHMAD, I.A. (1975). A nonparametric test for the monotonicity of a failure rate function. *Communications in Statistics – Theory & Methods*, **A4**, 967–974.

BARLOW, R.E. & DOKSUM, K. (1972). Isotonic tests for convex orderings. *Proceedings of the Sixth Berkeley Symposium*, **1**, 293–323.

BARLOW, R.E. & PROSCHAN, F. (1965). *Mathematical Theory of Reliability.* The SIAM Series in Applied Mathematics. John Wiley, New York.

BARLOW, R.E. & PROSCHAN, F. (1969). A note on tests for monotone failure rate based on incomplete data. *The Annals of Mathematical Statistics*, **40**, 595–600.

BARLOW, R.E. & PROSCHAN, F. (1975). *Statistical Theory of Reliability and Life testing Probability Models.* International Series in Decision Processes. Holt, Rinehart and Winston, Inc., New York.

BICKEL, P.J. (1969). Tests for monotone failure rate II. *The Annals of Mathematical Statistics*, **40**, 1250–1260.

BICKEL, P.J. & DOKSUM, K.A. (1969). Tests for monotone failure rate based on normalized spacings. *The Annals of Mathematical Statistics*, **40**, 1216–1235.

DÜMBGEN, L. (1999). Application of local rank tests to nonparametric regression. *Technical Report*, Medical University of Lübeck, Germany.

EPSTEIN, B. & SOBEL, M. (1953). Life testing. *Journal of the American Statistical Association*, **48**, 486–502.

GAIL, M.H. & GASTWIRTH, J.L. (1978a). A scale-free goodness-of-fit test for the exponential distribution based on the Lorentz curve. *Journal of the American Statistical Association*, **73**, 787–793.

GAIL, M.H. & GASTWIRTH, J.L. (1978b). A scale-free goodness-of-fit test for the exponential distribution based on the Gini statistic. *Journal of the Royal Statistical Society, Series B*, **40**, 350–357.

GERLACH, B. (1987). Testing exponentiality against increasing failure rate with randomly censored data. *Statistics*, **18**, 275–268.

GIJBELS, I., HALL, P., JONES, M.C. & KOCH, I. (2000). Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, **87**, 663–673.

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

HALL, P. & HECKMAN, N. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics*, **28**, 20–39.

HALL, P., HUANG, L.-S., GIFFORD, J. & GIJBELS, I. (2000). Nonparametric estimation of hazard rate under the constraint of monotonicity. *Journal of Computational and Graphical Statistics*, to appear.

HALL, P. & PRESNELL, B. (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B*, **61**, 259–277.

HUANG, J. & WELLNER, J. (1995). Estimation of a monotone density or monotone hazard under random censoring. *Scandinavian Journal of Statistics*, **22**, 3–33.

KLEFSJÖ, B. (1983). Some tests against aging based on the total time on test transform. *Communications in Statistics – Theory & Methods*, **A12**, 907–927

KUMAZAWA, Y. (1992). Tests for increasing failure rate with randomly censored data. *Statistics*, **23**, 17–25.

LAWLESS, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

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

MANN, H.B. (1945). Nonparametric tests against trend. *Econometrica*, **13**, 245–259.

MARSHALL, A.W. & PROSCHAN, F. (1965). Maximum likelihood estimation for distributions with monotone failure rate. *The Annals of Mathematical Statistics*, **36**, 69–77.

MUKERJEE, H. & WANG, J.-L. (1993). Nonparametric maximum likelihood estimation of an increasing hazard rate for uncertain cause-of-death data. *Scandinavian Journal of Statistics*, **20**, 17–33.

MYKYTYN, S. & SANTNER, T. (1981). Maximum likelihood estimation of the survival function based on censored data under hazard rate assumptions. *Communications in Statistics – Theory & Methods*, **A10**, 1369–1387.

NELSON, W.B. (1969). Hazard plotting for incomplete failure data. *Jour. Qual. Technology*, **1**, 27–52.

NELSON, W.B. (1970). Hazard plotting methods for analysis of life data with different failure modes. *Jour. Qual. Technology*, **2**, 126–149.

PADGETT, W.J. & WEI, L.J. (1980). Maximum likelihood estimation of a distribution function with increasing failure rate based on censored observations. *Biometrika*, **67**, 470–474.

PROSCHAN, F. & PYKE, R. (1967). Tests for monotone failure rate. *Fifth Berkeley Symposium*, **3**, 293–313.

ROBERTSON, T., WRIGHT, F.T. & DYKSTRA, R.L. (1988). *Order Restricted Statistical Inference*. New York: Wiley.

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

WANG, J.-L. (1986). Asymptotically minimax estimators for distributions with increasing failure rate. *The Annals of Statistics*, **14**, 1113–1131.

WATSON, G.S. & LEADBETTER, M.R. (1964a). Hazard Analysis I. *Biometrika*, **51**, 175–184.

WATSON, G.S. & LEADBETTER, M.R. (1964b). Hazard Analysis II. *Sankhyã, Series A*, **26**, 101–116.

		a_1			
		-0.2	-0.1	0.0	0.1
$\beta = 0$	global	0.001	0.017	0.05	0.189
	skiprest	0.010	0.025	0.05	0.095
	skipapprox	0.012	0.022	0.05	0.092
	skip	0.034	0.043	0.05	0.087
	local	0.029	0.043	0.05	0.089
$\beta = 0.3$	global	0.014	0.055	0.128	0.289
$\sigma = 0.2$	skiprest	0.066	0.118	0.184	0.267
	skipapprox	0.078	0.123	0.180	0.247
	skip	0.061	0.101	0.143	0.184
	local	0.063	0.087	0.142	0.190
$\beta = 0.3$	global	0.024	0.102	0.229	0.410
$\sigma = 0.1$	skiprest	0.139	0.255	0.336	0.478
	skipapprox	0.143	0.269	0.338	0.478
	skip	0.118	0.197	0.240	0.373
	local	0.147	0.219	0.301	0.393

Table 1. Proportion of times the null hypothesis is rejected at the 5% level for survival times with log hazard (4.1).

	(1.2): H_0: non-increasing hazard	(1.3): H_0: non-decreasing hazard
global	0.013	0.989
skiprest	0.148	0.651
skipapprox	0.198	0.738
skip	0.112	0.937
local	0.014	0.991

Table 2. The p-values for testing the ball bearings data.

		a_1			
		-0.2	-0.1	0.0	0.1
$\beta = 0$	global	0.008	0.032	0.095	0.230
	skiprest	0.016	0.041	0.064	0.117
$\beta = 0.3$	global	0.016	0.080	0.205	0.318
$\sigma = 0.2$	skiprest	0.042	0.105	0.185	0.262
$\beta = 0.3$	global	0.036	0.135	0.266	0.462
$\sigma = 0.1$	skiprest	0.103	0.194	0.285	0.390

Table 3. Proportion of times the null hypothesis is rejected at the 5% level for survival times with log hazard (4.1). Survival times are censored at a rate of 20%.

	(1.2): H_0: non-increasing hazard	(1.3): H_0: non-decreasing hazard
global	0.161	0.8655
skiprest	0.002	0.4735
skipapprox	0.003	0.533
skip	0.0085	0.3415
local	0.0005	0.3695

Table 4. The p-values for testing the electrical appliance data.

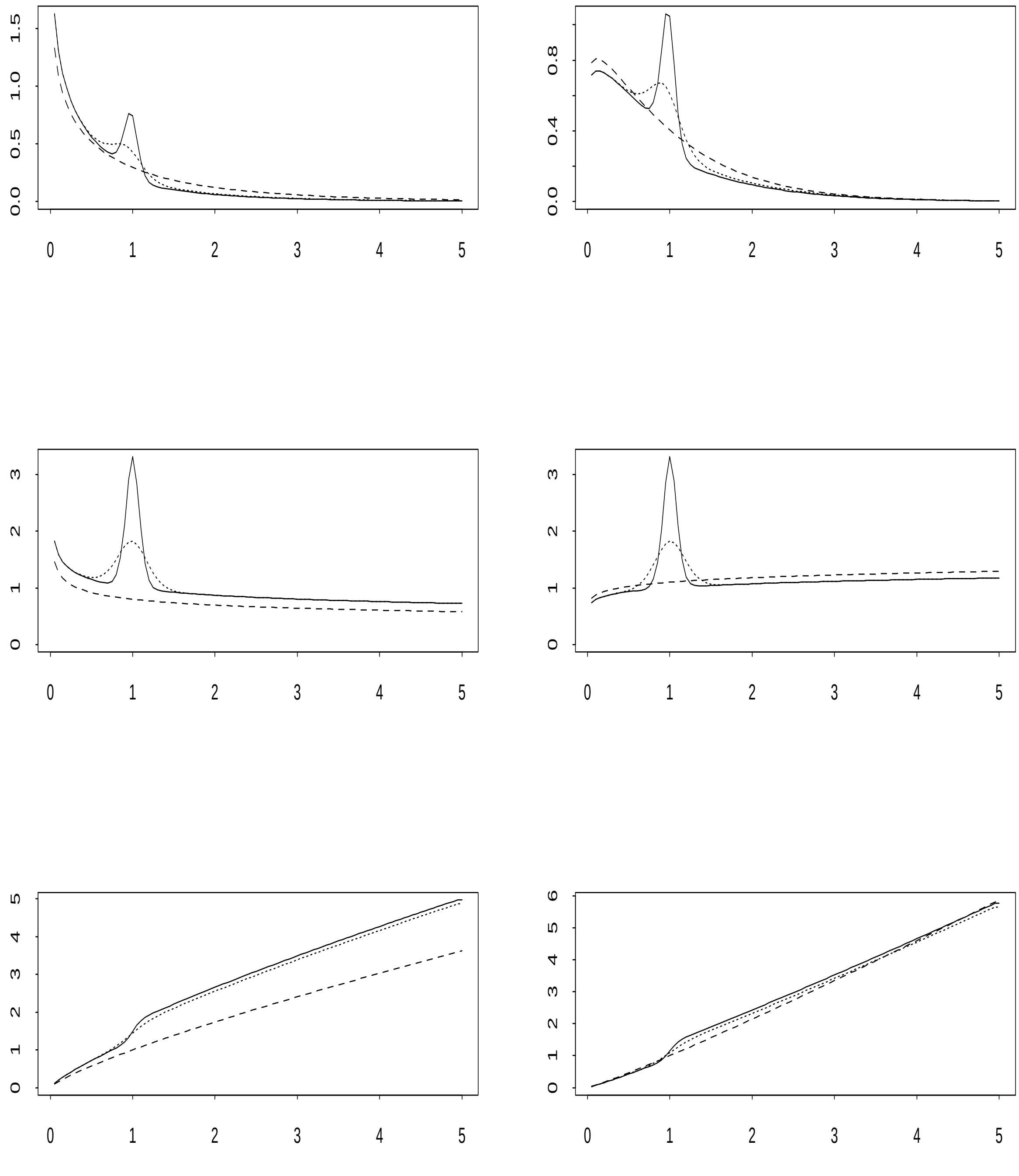

Figure 1. Densities (top row), hazards (middle row) and cumulative hazards (bottom row) for the log hazard (4.1) with $a_0 = 0$ and $\mu = 1$. The plots on the left side are with $a_1 = -0.2$ and the plots on the right side are with $a_1 = 0.1$. In each plot, the three curves correspond to $\beta = 0$ (dashed) $\sigma = 0.2$ (dotted) and $\sigma = 0.1$ (solid).

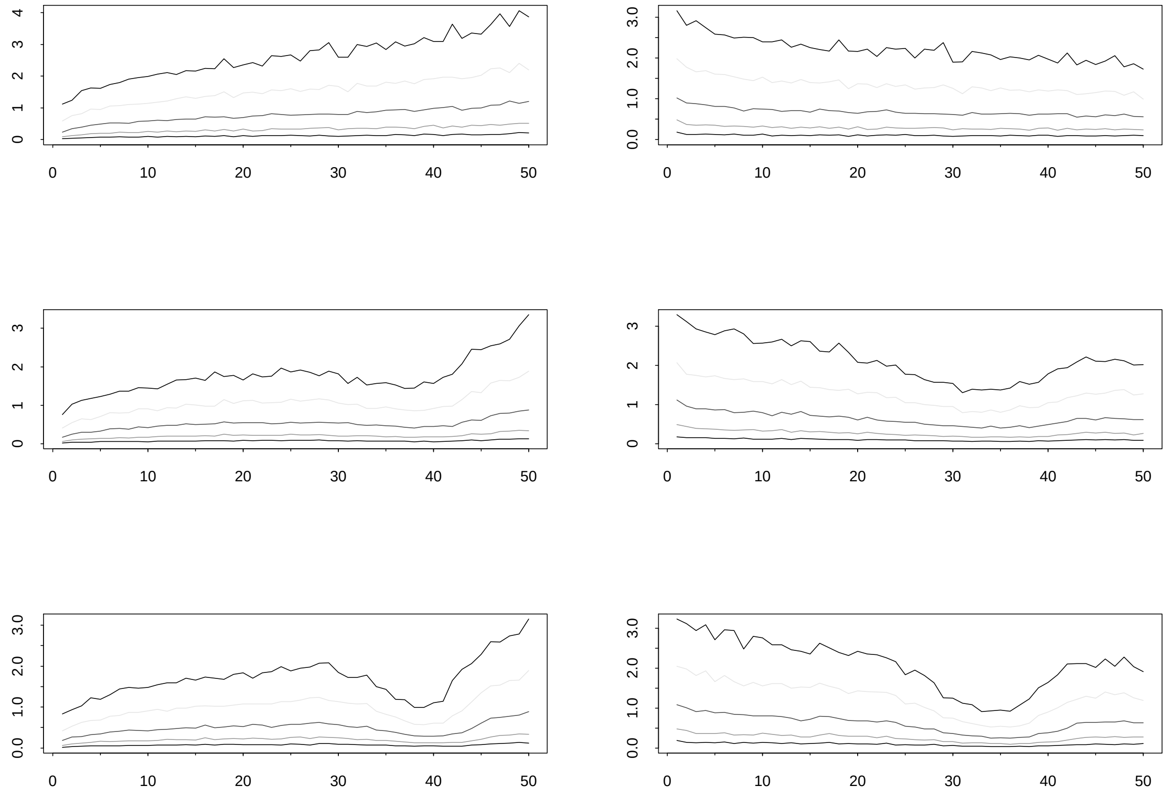

Figure 2. Plots of the pointwise quantiles (5th, 25th, 50th, 75th, 95th) of the normalized spacings $\bar{D}_{ni}$ versus i for $n = 50$. Data were generated using log hazard (4.1) with $a_0 = 0$ and $\mu = 1$. In the left column, we have taken $a_1 = -0.2$ and the right column $a_1 = 0.1$. In the first row, $\beta = 0$, so the lifetimes follow a Weibull distribution. In the second row $\sigma = 0.2$, and in the third row $\sigma = 0.1$.

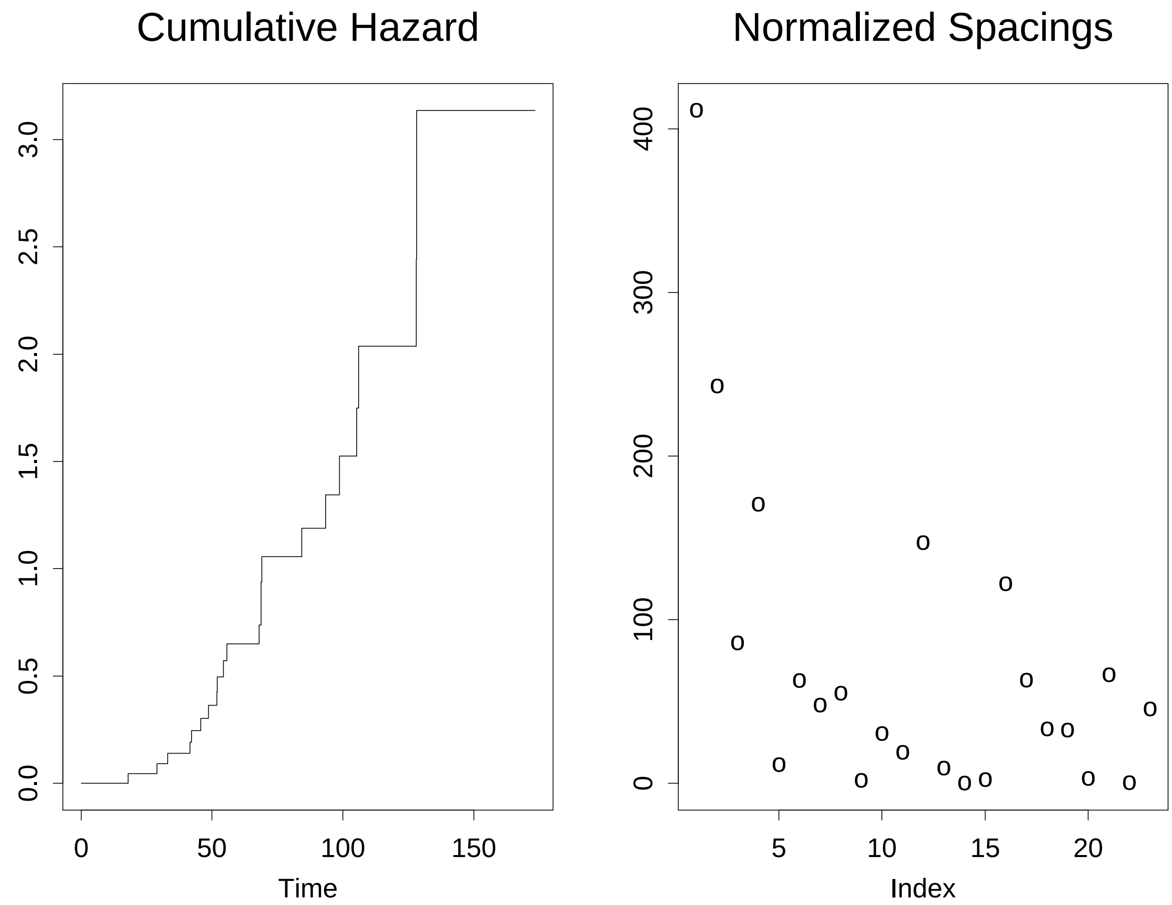

Figure 3. The bearings data. The left plot contains the estimate of the cumulative hazard. The right plot is of the normalized spacings $\bar{D}_{ni}$ versus i.

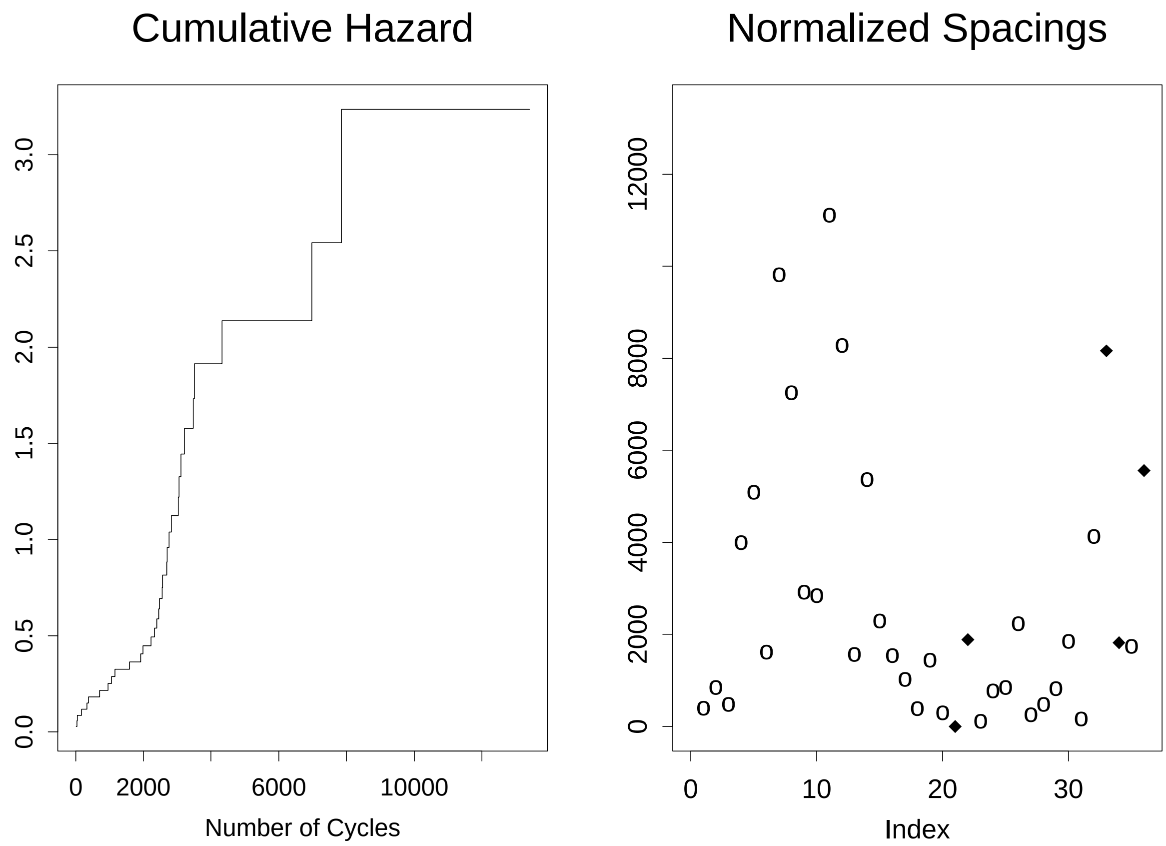

Figure 4. The electrical appliance data. The left plot contains the estimate of the cumulative hazard, based on the original data set. The right plot is of the normalized spacings $\bar{D}_{ni}$ versus i and is based on imputed data. Normalized spacings based on imputed data appear as diamonds.