Density and hazard rate estimation for right censored data using wavelet methods

Anestis Antoniadis, Gérard Grégoire
Laboratoire LMC - IMAG
Université Joseph Fourier
BP 53, 38041 Grenoble Cedex 09, France

and

Guy Nason
Department of Mathematics
University of Bristol
University Walk, Bristol
BS8 1TW, England

Keywords: SURVIVAL DATA; HAZARD RATE; WAVELET ESTIMATION

Abstract

This paper describes a wavelet method for the estimation of density and hazard rate functions from randomly right censored data. We adopt a nonparametric approach in assuming that the density and hazard rate have no specific parametric form. The method is based on dividing the time axis into a dyadic number of intervals and then counting the number of events within each interval. The number of events and the survival function of the observations are then separately smoothed over time via linear wavelet smoothers, and then the hazard rate function estimators are obtained by taking the ratio. We prove that the estimators possess pointwise and global mean square consistency, obtain the best possible asymptotic MISE convergence rate and are also asymptotically normally distributed. We also describe simulation experiments that show these estimators are reasonably reliable in practice. The method is illustrated with two real examples. The first uses survival time data for patients with liver metastases from a colorectal primary tumour without other distant metastases. The second is concerned with times of unemployment for women and the wavelet estimate, through its flexibility, provides a new and interesting interpretation.

1 Introduction

In life testing, medical follow up and other studies, the observation of the occurrence of the event of interest (called a failure) may be prevented by the previous occurrence of another event (called a censoring event). Such is the case for example in a clinical trial with fixed study time and random patient entry time, a clinical trial in which patients are lost to follow-up, or a clinical trial in which there are multiple causes of failure but interest centres on only one of them. In all these studies, interest focuses on estimating the underlying distribution density and hazard rate of the time to occurrence of the primary failure event.

In the above settings, popular methods for performing density and hazard rate estimation include kernel and nearest neighbour smoothing on the time axis (Beran 1981, Dabrowska 1987, Gray 1992). Penalized likelihood methods, such as those described by O'Sullivan (1988), Antoniadis (1989), Antoniadis and Grégoire (1990) have also been developed for hazard estimation. In all these methods, the programming to implement reasonably fast algorithms is not trivial. Kooperberg and Stone (1992) developed an approach based on MARS type adaptive regression spline models. A major limitation of their implementation is that their method tends to be computationally intensive.

Another traditional approach to density and hazard rate estimation is by orthogonal series (Kronmal and Tarter, 1968, Tanner and Wong, 1984). Recently, wavelet shrinkage curve estimation has become a well-known and mathematically sound technique for adaptively estimating functions. Optimal rates of convergence have been thoroughly examined for different observation schemes and L_2 loss by many authors. Software for fast wavelet smoothing is effectively implemented in many popular packages (see Nason and Silverman, 1994, and Buckheit and Donoho, 1995). Most current wavelet methods focus on density estimation or on ordinary regression (see e.g. Donoho and Johnstone, 1994, 1995, Donoho *et al.*, 1995, and Nason, 1996). This paper explores the possibility of applying an ordinary nonparametric wavelet smoother to the problem of estimating the density and hazard function of right censored data. The goal is to take advantage of fast wavelet methods and software for nonparametric regression and to simplify the task of implementing software for the more complex problem of hazard smoothing.

To briefly describe the proposed method, let $X_1, X_2, \dots, X_n$ denote lifetimes (times to failure) for the n subjects under study, and let $C_1, C_2, \dots, C_n$ be the corresponding censoring times. The *observed* random variables are then Z_i and δ_i where

$$Z_i = \min(X_i, C_i) \quad \text{and} \quad \delta_i = \mathbf{I}_{[X_i \le C_i]},$$

where $\mathbf{I}_A$ denotes the indicator of event A. So $\delta_i = 1$ indicates that the i^{th}

subject's observed time is not censored. We assume throughout this paper that:

(i) $X_1, X_2, \ldots, X_n$ are nonnegative and i.i.d. with common continuous cdf F and continuous density f,

(ii) $C_1, C_2, \ldots, C_n$ are nonnegative and i.i.d. with common continuous cdf G and continuous density g,

(iii) lifetimes and censoring times are independent.

The hazard rate function is given by

$$\lambda(t) = \frac{f(t)}{1 - F(t)}, \quad F(t) < 1.$$

In the censored case if $G(t) < 1$, we have

$$\lambda(t) = \frac{f(t)(1 - G(t))}{(1 - F(t))(1 - G(t))}, \quad F(t) < 1. \tag{1}$$

If we let

$$L(t) = \mathbb{P}\{Z_i \leq t\},$$

we see that

$$1 - L(t) = (1 - F(t))(1 - G(t)). \tag{2}$$

Let

$$f^*(t) = f(t)(1 - G(t)), \tag{3}$$

be the density of those observations still to fail and substitute (2) and (3) into (1) to obtain

$$\lambda(t) = \frac{f^*(t)}{1 - L(t)}, \quad L(t) < 1. \tag{4}$$

Expression (4) is the key formula underlying our method. We will first estimate the subdensity of the observed failures $f^*(t)$. Dividing this estimate by an appropriate estimator of the probability $1 - L(t)$ of follow-up continuing to t, will provide an estimate of the hazard rate function $\lambda(t)$.

To estimate the subdensity f^*, divide the time axis into a dyadic number of small intervals, count the number of events (0 or 1) for each case in each interval and divide these by the interval widths. Treating the contributions of each case to each interval as if they were independent observations, regress the number of events on t using a nonparametric wavelet smoother to obtain an estimate of the subdensity. We can summarize our subdensity estimation procedure by the following steps:

Step 1 Choose $\Delta > 0$ and bin the observed failures into $K+1$ bins of length Δ.

Step 2 Use a wavelet regression estimate on the binned data to obtain an estimate of the subdensity, i.e. choose a resolution $j(n)$, compute the wavelet coefficients of the binned data at scale $j(n)$ and take the resultant wavelet transform as a smooth estimate of f^*.

An estimator of the hazard rate function will then be formed by dividing the subdensity estimate by an empirical wavelet estimator of the probability of remaining at risk.

An immediate advantage of our method over other methods is its speed and its ease of computation. Wavelet estimators can be computed through an $\mathcal{O}(n)$ algorithm, i.e. above step 2 requires $\mathcal{O}(K)$ computations. The binning in step 1 can be done particularly fast and simply. Hence, the number of evaluations necessary to compute our subdensity estimator certainly is linear in the number of observations n.

It is perhaps worth pointing out that our approach differs from other orthogonal series approaches to density estimation, since the binning process allows us to adapt nonparametric wavelet *regression* methods to estimation of a density. Existing wavelet density estimation methods such as those developed by Donoho *et al.* (1995) are based on thresholding empirical wavelet coefficients. For moderate sample sizes the resultant density estimates can have a very rough appearance (see Tribouley, 1995).

The general formulation and some theoretical properties of the method are discussed in the following section. Simulation results are given in Section 3, and some real examples are given in Section 4.

2 Estimation of the hazard rate function

2.1 Notation and model set-up

Before introducing our estimator we need some preliminaries. Define

$$T_F = \sup\{t : F(t) < 1\}$$

$$T_G = \sup\{t : G(t) < 1\}$$

$$T_L = \sup\{t : L(t) < 1\} = \min\{T_F, T_G\}.$$

Estimates will only be computed over a finite interval $[0, \tau]$ where τ is assumed to satisfy $\tau < T_L$. Note that, if $Z_{(i)}$ denotes the ith order statistic of the sequence Z_i, it is easy to show that $T_{L_n} = Z_{(n)} \to T_L$ *a.s* as $n \to \infty$ (see Carbonez *et al.*, 1995). So in practice we take $\tau = Z_{(n)}$.

Let N be an integer that may depend on n and define a dyadic grid (evaluation points)

$$t_k = \frac{k\tau}{2^N}; \quad k = 0, \ldots, K = 2^N - 1,$$

with $\Delta = \tau 2^{-N}$ the interpoint distance on the grid. The time axis is the interval $[0, \tau]$ divided into $K+1$ intervals of length Δ centred on t_k with endpoints

$$\tau_0 = -\frac{\Delta}{2}; \quad \tau_k = t_k - \frac{\Delta}{2}, \; k = 1, \ldots, K; \tau_{K+1} = \tau.$$

The kth interval is denoted by J_k: so $J_k = [\tau_k, \tau_{k+1}[$, $k = 0, \ldots, K-1$ and $J_K = [\tau_K, \tau]$. Now using the observations, a data set of $(K+1)n$ records is created, consisting of (Y_{ik}, t_k) where

$$Y_{ik} = \mathbf{I}_{J_k}(Z_i)\delta_i \quad i = 1, \ldots, n; k = 0, \ldots, K,$$

is the indicator that a noncensored event for subject i falls within the time interval J_k. Finally, let U_k be the proportion of failures observed to fail in the interval J_k, in other words:

$$U_k = \frac{1}{n}\sum_{i=1}^{n} Y_{ik}, \quad k = 0, \ldots, K.$$

Note that U_k/Δ are crude estimators of the subdensity values $f^*(t_k)$, defining a histogram-type estimator at $K+1$, a power of 2, dyadic points. These will be further smoothed by a discrete fast wavelet method. Computationally, such a histogram method has the advantage that it reduces the number of points for smoothing from n to $K+1$. Such binning to improve speed in smoothing for the i.i.d. case has been discussed for example by Härdle (1991), Härdle and Scott (1992).

Later we shall prove large sample properties for our estimator. In preparation for this we first investigate the moments of the variables U_k.

Lemma 1
Assume that the subdensity f^* is continuously differentiable on $[0, \tau]$. Then, as $\Delta \to 0$ for $n \to \infty$ we have:

$$I\!E[U_k/\Delta] = f^*(t_k) + \mathcal{O}(\Delta), \tag{5}$$

$$Var[U_k/\Delta] = \frac{f^*(t_k)}{n\Delta} + \mathcal{O}(n^{-1}), \tag{6}$$

and

$$Cov[U_k/\Delta, U_\ell/\Delta] = -\frac{1}{n} f^*(t_k) f^*(t_\ell) + \mathcal{O}\left(\frac{\Delta}{n}\right), \qquad k \neq \ell. \tag{7}$$

The assertions of Lemma 1 are fairly minor modifications of standard results for kernel density estimates and their proof is therefore omitted.

2.2 Estimating the subdensity f^* of observed failures

To obtain our estimate of f^* we will smooth the binned (equally spaced) data U_k/Δ via an appropriate linear wavelet smoother. An immediate advantage of our approach over other methods is its speed and its ease of computation. Wavelet estimators can be computed using an $\mathcal{O}(n)$ algorithm, i.e. the above smoothing step only requires $\mathcal{O}(K)$ computations. The binning process can be performed particularly quickly and simply. Hence, the number of evaluations necessary to compute our estimator is certainly linear in the number of observations n which is much faster than kernel or logspline estimation algorithms.

Our wavelet estimator will be based on wavelets on the interval because our problem is confined to an interval. Wavelets and multiresolution analyses of $L^2([0,\tau])$ have been introduced and explicitly constructed recently by Cohen *et al.* (1993). Their construction uses "interior" and "edge" orthonormal scaling functions at every resolution, so that the total number is exactly 2^j at resolution j.

The idea underlying such an approach is to express any function $f \in L^2([0,\tau])$, with appropriate allowance for end effects, in the form

$$f(t) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0,k}\phi_{j_0,k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^{j_0}-1} \beta_{j,k}\psi_{j,k}(t)$$

for collections of functions $\{\phi_{j_0,k}\}$ and $\{\psi_{j,k}\}$ which form an orthogonal basis for $L^2([0,\tau])$. The $\phi_{j_0,k}$ are defined by

$$\phi_{j_0,k}(x) = 2^{j_0/2}\phi(2^{j_0}x - k)$$

for some function ϕ with $\int \phi(x)\,dx = 1$ called the *scaling function*; these allow approximation of f at resolution j_0. The $\{\psi_{j,k}\}$ are generated in a similar way from a *mother wavelet* ψ, and represent the detail in f at resolutions finer than j_0.

The edge scaling functions are adapted in such a way that all the polynomials on $[0,\tau]$ of degree less than an integer, depending on the number of vanishing moments of the generating scaling function, can be written as linear combinations of the scaling functions at any fixed scale. As on the whole real line, no explicit analytic expression for the scaling functions and for the wavelets is available. But for practical applications all that is really needed are the filter coefficients and the multiresolution framework. A list of suitable coefficients is available in the paper of Cohen *et al.* (1993).

From now on, ϕ and ψ will denote the scaling function and the mother wavelet on $[0,\tau]$ associated with an r-regular multiresolution analysis of $L^2([0,\tau])$. We recall that a multiresolution analysis is said to be r-regular

($r \geq 0$) if ϕ is an element of the Hölder space C^r, and if both ϕ and its derivatives have a fast decay,

$$|\partial^\alpha \phi(x)| \leq C_m (1+|x|)^{-m}, \quad \forall m \in I\!N, \quad 0 \leq \alpha \leq r,$$

for some sequence of finite constants C_1, C_2, ... One can prove that if a multiresolution analysis is r-regular, the wavelet ψ is also in C^r and has vanishing moments up to the order r (see e.g. Daubechies (1992), Corollary 5.2)

$$\int_0^\tau x^k \psi(x) dx = 0 \quad \text{for} \quad 0 \leq k \leq r.$$

The converse is generally false, and the number of vanishing moments is usually larger than the regularity of the multiresolution analysis. The smoother wavelets provide not only orthonormal bases for $L^2([0,\tau])$, but also unconditional bases for function spaces consisting of more regular functions.

An advantage of having a high number of vanishing moments for ψ is that the fine scale wavelet coefficients of a function are essentially zero where the function is smooth. Since $\int \phi(x) dx = 1$, the same thing can never happen for the $< f, \phi_{j,k} >$, but it is possible to construct compactly supported orthonormal wavelets such that the scaling function ϕ has L vanishing moments, i.e.

$$\begin{aligned}
\int_0^\tau \phi(x)\, dx &= 1\,, \\
\int_0^\tau x^k \phi(x)\, dx &= 0 \quad 1 \leq k \leq L\,, \\
\int_0^\tau x^k \psi(x)\, dx &= 0 \quad 0 \leq k \leq L\,.
\end{aligned}$$

Such wavelets were constructed by Daubechies (1992) and were named *coiflets* after Ronald Coifman who asked for their construction.

We will assume hereafter that the scaling function ϕ is a coiflet of order $L = 2q$ with $L > m+1$, where m is the assumed order of differentiability of f. Let V_j and W_j be the approximation and detail spaces associated with the multiresolution analysis of $L^2([0,\tau])$ generated by ϕ. Since for some integer j_0

$$L^2([0,\tau]) = V_{j_0} \oplus (\oplus_{j \geq j_0} W_j)\,,$$

the function f^* admits the following generalized Fourier expansion in L^2:

$$f^*(t) = \sum_{k=0}^{2^{j_0}-1} < f^*, \phi_{j_0,k} > \phi_{j_0,k}(t) + \sum_{j \geq j_0} \sum_{\ell=0}^{2^j-1} < f^*, \psi_{j,\ell} > \psi_{j,\ell}(t),$$

with $< f, g >$ defined by $\int_0^\tau f(t) g(t)\, dt$.

Finally let us mention some approximation properties of regular wavelets that will be used in the sequel of this paper (for a detailed account see Mallat (1989)).
If f belongs to the Sobolev space $H^{\delta}([0,\tau])$ and the multiresolution analysis is r-regular, then

$$\|f - P_j f\|_{L^2} \leq o(2^{-j\min(\delta,r)}) \quad \text{as } j \to \infty\,, \tag{8}$$

where $P_j f$ denotes the projection of f onto the approximation space V_j.

The nested structure of a multiresolution analysis leads to an efficient tree-structured algorithm for the decomposition of functions in V_N for which the fine-scale theoretical wavelet coefficients $< f^*, \phi_{N,k} >$ are given. However, when a function is given in sampled form, one typically does not have access to the fine scale integrals $< f^*, \phi_{N,k} >$, which are needed to initialize the fast wavelet transform. In applications it is widely assumed that the $< f^*, \phi_{N,k} > \simeq 2^{-N/2} f^*(k/2^N)$, but such an approximation is rarely justified. When ϕ is a regular enough coiflet, the accuracy of this approximation can be controlled by the following lemma (the proof may be found in Antoniadis, 1996).

Lemma 2
Given an m-differentiable function r, let $r^{\{j\}}(k) = 2^{j/2} < r, \phi_{j,k} >$. With $L > m+1$, the following uniform (in k; $0 \leq k \leq 2^j$) bound holds:

$$\left| r^{\{j\}}(k) - r\left(\frac{k}{2^j}\right)\right| \leq C_1\, 2^{-jm},$$

where C_1 is a constant only depending on the coiflet ϕ and its support length.

By Lemma 2 one is therefore able to approximate the coefficients $< f^*, \phi_{N,k} >$ by $2^{-N/2} f^*(t_k)$, $0 \leq k \leq 2^N - 1$, with an error $\mathcal{O}(2^{-N/2} 2^{-Nm})$. Therefore, a reasonable estimate of the projection $\Pi_N f^*$ of f^* onto the finest available scale N is

$$\tilde{f}^*_N(t) = 2^{-N/2} \sum_{k=0}^{K} \frac{U_k}{\Delta} \phi_{N,k}(t) \tag{9}$$

where $\phi_{N,k}(t) = 2^{N/2}\phi(2^N t - k)$ are the finest scale scaling functions. This follows by observing that $I\!E(U_k/\Delta) \simeq f^*(t_k)$. More precisely, we have:

Lemma 3
Assume that the subdensity f^* is m-times continuously differentiable on $[0,\tau]$. Then, as $\Delta \to 0$ for $n \to \infty$ we have:

$$I\!E[\tilde{f}^*_N(t)] = \Pi_N f^*(t) + \mathcal{O}(\Delta) \tag{10}$$

and

$$\mathrm{Var}[\tilde{f}^*_N(t)] = \mathcal{O}\{(n\Delta)^{-1}\} + \mathcal{O}(n^{-1}). \tag{11}$$

Proof. The proof may be found in the appendix.

The above calculations suggest that the observed binned values U_k, are equivalent to a "raw" estimator $\tilde{f}_N^*$, which now lies in the Sobolev space $H^m([0,\tau])$ by the m-regularity of ϕ. Combining the results of Lemma 3 gives the point-wise consistency of the estimator $\tilde{f}_N^*(t)$. This estimator, while presenting a very small bias on $[0,\tau]$, leads to an oscillatory solution almost interpolating the binned data. It is easy to see that the best convergence rate for the mean integrated squared error (MISE) for $\tilde{f}_N^*(t)$ is $\mathcal{O}(n^{-2/3})$ obtained by choosing $\Delta = n^{-1/3}$. In order to smooth the data with a better rate, we will associate with each sample size n a resolution $j(n) < N$ and estimate the unknown function f^* by

$$\hat{f}_n = P_{V_{j(n)}} \tilde{f}_N^*, \tag{12}$$

the orthogonal projection of $\tilde{f}_N^*$ onto the "smoother" approximation space $V_{j(n)}$. The parameter $j(n)$ governs the smoothness of our estimator. It is important to choose it judiciously because it controls the trade-off between fidelity to the data and the smoothness of the resulting solution. Too small a value of $j(n)$ leads to an over-smoothed, biased solution. From a theoretical viewpoint, in the derivation of asymptotic results, the smoothing parameter must tend to infinity at the correct rate as the amount of information in the data grows to infinity. The following theorem addresses some of the asymptotic properties of $\hat{f}_n$, defined by equation (12).

Theorem 1
Under the assumptions imposed on f^* in this section, the MISE of $\hat{f}_n$ defined by

$$R_n = \mathbb{E}\left[\int_0^\tau \left\{\hat{f}_n(t) - f^*(t)\right\}^2 dt\right]$$

satisfies, as $n \to \infty$, $\Delta \to 0$ and $j(n) \to \infty$ with $n2^{-j(n)} \to \infty$,

$$R_n \leq \mathcal{O}(2^{-2j(n)m}) + \mathcal{O}(\Delta^2) + \mathcal{O}\left(\frac{2^{j(n)}}{n}\right).$$

The first term $\mathcal{O}(2^{-2j(n)m}) + \mathcal{O}(\Delta^2)$ in the upper bound of the risk R_n corresponds to an upper bound on the squared integrated bias of the estimator $\hat{f}_n$, while the second term $\mathcal{O}(\frac{2^{j(n)}}{n})$ is an upper bound on its variance. From these expressions it is easily seen that a small value of $j(n)$ causes large bias, whilst a large value of $j(n)$ can cause large variance.
Proof. The proof appears in the appendix.

Note that according to Theorem 1, an optimal choice for $j(n)$ and N (or Δ) is $j(n) = \frac{1}{2m+1}\log_2(n)$ and $N \geq \frac{m}{2m+1}\log_2(n)$. In Ibragimov and Khasminski (1982), Stone (1982) and Nussbaum (1985) it has been proved that the best global convergence rate in the MISE sense of *any* nonparametric estimator of a density in the class of m-smooth functions that we consider ($m \geq 1$), is $\mathcal{O}(n^{-u})$ with $u = \frac{2m}{2m+1}$. It is clear that, with regular enough coiflets, our estimator asymptotically attains this best possible convergence rate.

The asymptotic normality of $\hat{f}_n(t)$ for any dyadic point of the interval $[0, \tau]$ of the form $k/2^q$ with q integer, is shown by the following theorem.

Theorem 2
Under the same assumptions as in Theorem 1, and assume that for $n \to \infty$ $n\Delta \to \infty$, $n\Delta^3 \to 0$ and $n\Delta 2^{-j(n)(2m-1)} \to 0$. Then for any dyadic point $t \in [0, \tau]$ $\sqrt{n\Delta}\left(\hat{f}_n(t) - f^*(t)\right)$ converges in distribution to a zero mean Gaussian random variable with variance $f^*(t)w^2$ where $w^2 = \sum_{k=-\infty}^{k=+\infty} \phi^2(k)$.

Proof. The proof appears in the appendix.

2.3 Estimating the probability $L(t)$ of follow-up continuing to t

In order to be able to estimate the hazard rate function we need an appropriate and consistent estimator $\hat{L}_n(t)$ of the cumulative distribution function $L(t)$.

Given the set of i.i.d. observations $Z_1, \ldots, Z_n$ from the common distribution function L, the standard nonparametric estimator of L is the empirical distribution function L_n defined by

$$L_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{[Z_i \leq t]}.$$

This estimator L_n of L was proved by Dvoretsky, Kiefer and Wolfowitz (1956) to be asymptotically minimax among the collection of all distribution functions. Therefore, in the absence of additional information about the shape of L the empirical distribution function L_n is the optimal estimator for L in the asymptotically minimax sense. Although L_n is asymptotically optimal, it does not take fully into account the smoothness of L (i.e., the existence of a density for L). It therefore seems reasonable to consider a continuous estimator of L which is better adapted to this situation. Thus, in relatively general situations, an estimator of the form

$$\hat{L}_n(t) = \frac{1}{n}\sum_{i=1}^{n} H_n(t - Z_i), \quad t \in [0, \tau] \tag{13}$$

where $\{H_n\}$ is a sequence of continuous cdfs required to converge weakly to cdf of the delta distribution centred at 0 has been suggested in the literature (see Winter, 1973, 1979, Puri and Ralescu, 1986, Yukish, 1989). Such estimators arise quite naturally as integrals of density estimators of the kernel type. The estimator that we are going to use in this paper will be defined by

$$\hat{L}_n(t) = \int_0^t \hat{\ell}_n(x)\,dx, \quad t \in [0, \tau]$$

where $\hat{\ell}_n$ is a traditional histogram type estimator of the density ℓ of L. This estimator of L is nothing else than an integrated Haar transform of the data and may be viewed as a wavelet estimator of the survival function. The interest in such an estimator is primarily due to its simple structure as averages over independent random variables. The question of whether using a different smoother for $\hat{L}_n$ could give an improved hazard rate function estimator has not been investigated.

Let $\phi(t) = \mathbf{I}_{[0,\tau]}(t)$ be the indicator function of $[0, \tau]$ and denote by $\phi_{j,k}$ the translated and dilated functions

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k).$$

It is not difficult to see that, as j tends to ∞, $2^{j/2}\phi_{j,0}(t)$ converges to the delta distribution centred at 0, and therefore $\Phi_{j,0}(t) = 2^{j/2}\int_0^t \phi_{j,0}(x)dx$ is a sequence of continuous cdf converging weakly to the cdf of the delta distribution centred at 0. Let $\tilde{j}(n)$ be a sequence of scales such that $\tilde{j}(n) \to \infty$ as $n \to \infty$. Put $\hat{\ell}_n(t) = \frac{1}{n}\sum_{i=1}^n 2^{\tilde{j}(n)/2}\phi_{\tilde{j}(n),0}(t - Z_i)$ and define $\hat{L}_n(t) = \frac{1}{n}\sum_{i=1}^n \Phi_{\tilde{j}(n),0}(t - Z_i)$. Since by our assumptions L is continuously differentiable on $[0, \tau]$, the results of section 3 in Winter (1979) apply. More precisely we have

Proposition 1 (Winter (1979))
Suppose that $\lim_{n\to\infty} 2^{-\tilde{j}(n)}\{n/\log\log n\}^{1/2} = 0$. *Then*

$$\sup_{t\in[0,\tau]} |\mathbb{E}\{\hat{L}_n(t)\} - L(t)| = o(\{2n/\log\log n\}^{-1/2})$$

$$\sup_{t\in[0,\tau]} Var(\hat{L}_n(t)) \le \sup_{t\in[0,\tau]} Var(L_n(t))$$

and $\hat{L}_n$ *has the Chung-Smirnov property, that is*

$$\limsup_{n\to\infty}\left(\{2n/\log\log n\}^{1/2}\sup_{t\in[0,\tau]}|\hat{L}_n(t) - L(t)|\right) \le 1, \quad a.s.$$

The first assertion of Proposition 1 follows from the proof of Theorem 3.3 in Winter (1979). The second assertion is a consequence of his Lemma 2.3 (b) and the Chung-Smirnov property is given by his Theorem 3.3.

If the resolution $\tilde{j}(n)$ is chosen to be N, the log of the number of bins that we have used for estimating the subdensity f^*, then it is easy to see that the computation of $\hat{L}_n$ can be performed in $\mathcal{O}(N)$ operations. Note also that under the assumptions of Proposition 1, our estimate $\hat{L}_n$ is mean square consistent with a rate $\mathcal{O}(\{2n/\log\log n\}^{-1})$.

2.4 Estimating the hazard rate

Our estimator of $\lambda(t)$ is defined to be

$$\hat{\lambda}_n(t) = \frac{\hat{f}_n(t)}{1 - \hat{L}_n(t)},$$

where $\hat{L}_n(t)$ has been defined in the previous section. Our purpose here is to study the large sample properties for $\hat{\lambda}_n(t)$. More precisely we have the following Proposition whose proof is given in the appendix:

Proposition 2
Under the assumptions made in this section, if $j(n) = \frac{1}{2m+1}\log_2(n)$ *and* $N \geq \frac{m}{2m+1}\log_2(n)$, *then as* $n \to \infty$, *and for any* $t \in [0, \tau]$,

$$\mathbb{E}[(\hat{\lambda}_n(t) - \lambda(t))^2] = \mathcal{O}(n^{-2m/(2m+1)}).$$

Proof. The proof is given in the appendix (recall also that $\tau < T_L$).

The weak consistency of the hazard rate function estimator is a direct consequence of the above proposition. Moreover, since both $\hat{f}_n$ and $\hat{L}_n$ are consistent and because $\hat{\lambda}$ is a continuous function of these two quantities in $[0, \tau]$, from the proof of Proposition 2 it follows that the asymptotic distribution and the MISE of $\hat{\lambda}_n$ are the same as those of

$$\left[\frac{\hat{f}_n(t) - f^*(t)}{1 - L(t)} - \frac{f^*(t)}{(1 - L(t))^2}[L(t) - \hat{L}_n(t)]\right].$$

Since f^* is continuous on $[0, \tau]$ and since $(1 - L(t))$ is uniformly bounded away from 0 on $[0, \tau]$, and because the MISE of $\hat{L}_n$ is asymptotically smaller than that of $\hat{f}_n$, it follows, by the asymptotic normality of $\hat{f}_n$, that for any dyadic point in $[0, \tau]$, $\hat{\lambda}_n(t)$ is asymptotically normal. Moreover the asymptotic rate of the MISE of $\hat{\lambda}_n$ is the same as that of $\hat{f}_n$ which is given by Theorem 1.

3 Simulations

An advantage of using simulated data for examples involving censoring is that one knows not only the true density function from which the data were generated, but also the actual values of the sample data before the censoring took place. All simulations were run in S, using the built-in random number generators and the `WaveThresh` package of Nason. A well-known shortcoming of any orthogonal series estimator (including wavelets) is that except for rather special smoothers they are not guaranteed to be nonnegative. In the simulations $\hat{f}_n(t)$ and $\hat{L}_n(t)$ were replaced by

$$\hat{f}_+(t) = \max(\hat{f}_n(t), 0) \quad \text{and} \quad \hat{L}_+(t) = \max(\hat{L}_n(t), 0).$$

Such a procedure defines estimates that have a smaller mean squared errors than the original ones (see Efromovitch, 1989). Moreover, since the hazard estimates are very unstable and have little meaning when few subjects were left at risk, hazard estimates were only computed at points where $L(t) > 0.5$. Indeed, one should be wary about extrapolation beyond the range of the data. In particular when all observations beyond a certain point are censored conclusions about the right tail of the density may be unreliable. In binning the data interval lengths were chosen according to the rules suggested by the theorems of the previous section, that is by taking $\Delta = n^{-1/2}$. The resolution $j(n)$ was chosen by folded cross-validation (see Nason 1996).

3.1 Simulation 1

Figure 1 shows the wavelet estimates for the observed failure subdensities and the hazard rate functions for a traditional censoring scheme.

We generated a sample Y_i, $1 \le i \le n$, from the Gamma distribution with shape parameter 5 and scale parameter 1 and an independent sample C_i, $1 \le i \le n$, from the exponential distribution with mean 6 (the mean was chosen so as to yield about 50% censoring). The results reported in this subsection are based on samples of size $n = 200$. In binning the data, we used an interval length $\Delta = 0.07$. The resolution $j(n)$ was obtained by folded cross validation with an average value of 2 across simulations. Scaling functions corresponding to coiflets with four zero moments were used.

The solid line in the left panel of Figure 1 is the subdensity estimate $\hat{f}_1$ based on A_i, where $A_i = X_i$ if $X_i \le C_i$. The dotted line represents the true underlying subdensity f_1. Similarly, the right panel of Figure 1 displays the corresponding estimated hazard rate function $\hat{\lambda}_1$.

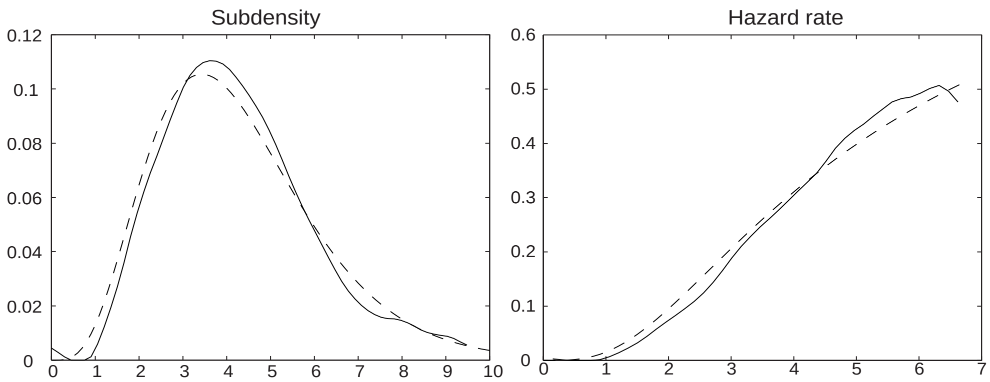

Figure 1: Wavelet estimate for the subdensity of observed failures for a traditional censoring scheme described in section 3.1. The solid line is the density estimate based on the actual data of size $n = 200$ with 50% censoring) and the dotted line is the true Gamma density. The right panel displays the wavelet estimate (solid line) for the observed failure hazard rate for the same data with the dotted line representing the true hazard rate.

3.2 Simulation 2

The censoring scheme for Figure 2 is the same as the one for Figure 1. Here the X_is were generated from the bimodal density f_2 that was used in Kooperberg and Stone (1992):

$$f_2 = 0.8g + 0.2h,$$

where g is the (lognormal) density of $\exp(Z/2)$, with Z having the standard normal distribution, and h is the normal density with mean 2 and standard deviation 0.17. The C_is were generated from the exponential distribution with mean 2.5. The sample size of the simulated data is $n = 200$. The solid lines in Figure 2 represent the estimates. The dotted line represents the true density f_2 (left panel) and hazard rate function λ_2 (right panel).

From Figure 1 we observe that it is possible in practice to recover well the underlying density of interest from information that is available in studies with right censored data. Even for the bimodal density, our method of density estimation does a decent job. For sample sizes of the order $n = 100$, however, the estimate for the height of the second mode was not very accurate. Further examination suggested that this appears to be caused primarily by sampling variation (the number of data points close to the second mode, ignoring censoring, is binomial with parameters $n = 100$ and $p = 0.2$). Although the total percentage of censoring is typically less than 40%, about 55% of the cases in the range of the second mode get censored. Nevertheless, when the sample size is large this censoring has almost no influence on the fit.

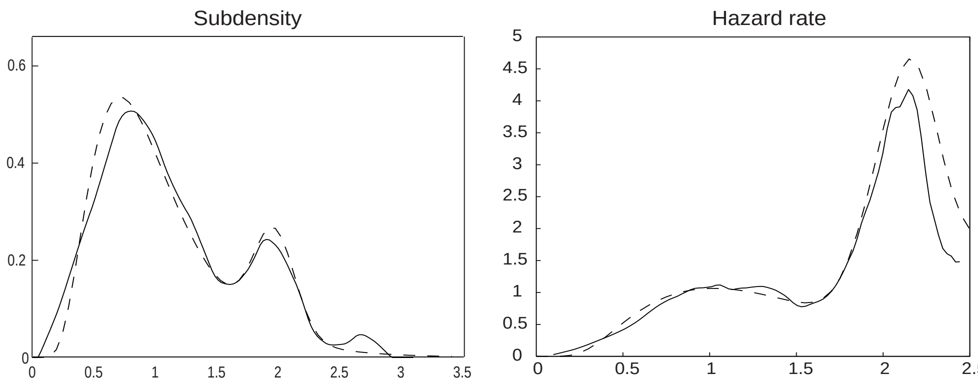

Figure 2: Wavelet estimate for the bimodal subdensity of observed failures described in section 3.2. The solid line is the subdensity estimates based on the actual data of size $n = 200$ and the dotted line is the true bimodal subdensity. The right panel displays the wavelet estimate (solid line) for the observed failure hazard rate for the same data with the dotted line representing the true hazard rate.

3.3 Mean-squared errors

To end this section, we summarize some simulation results on the mean square errors of our estimators. The first set of simulations examined the average mean squared error of the subdensity estimator as a function of the size Δ, estimating the optimal $j(n)$ by folded cross-validation.

Samples of size $n = 200$ and $n = 500$ were generated from the two different true hazard functions whose subdensities were shown in Figures 1 and 2.

The estimators used $K = 16$; 32; 64. The average mean squared errors AMSE(f^*) and AMSE(λ) were estimated by averaging the average squared errors

$$ASE(f^*) = K^{-1} \sum_k (\hat{f}_n(t_k) - f^*(t_k))^2 \quad ASE(\lambda) = K^{-1} \sum_k (\hat{\lambda}_n(t_k) - \lambda(t_k))^2$$

for each sample and averaging over the samples. Because the mean squared error of the estimator can be especially important at large t with small risk sets, this risk (denoted by AMSE2) was also calculated restricting the sum over time to points with $t_k \leq 6$ for the density f_1 and $t_k \leq 2.5$ for the bimodal density f_2. The results on the estimated subdensities from a total of 200 repetitions for each of the 2 models are given in Table 1 for each sample size $n = 200$ and $n = 500$.

Table 2 summarizes the results for the hazard rate estimates on the same simulated samples used to obtain the results of table 1.

Pointwise MSEs for the subdensity and the hazard rates were also tabulated at the points $t = 0.5$, $t = 0.75$, $t = 1.5$ and $t = 3$. Not surprisingly,

Table 1: Average mean squared errors ($\times 10^{-5}$) for subdensity estimation based on 200 repetitions of the simulations given in section 3.1 and 3.2 for sample sizes $n = 200$ and $n = 500$.

		Subdensity f_1		Subdensity f_2	
	Bins K	$n = 200$	$n = 500$	$n = 200$	$n = 500$
AMSE	16	25.7	17.6	671	603
	32	18.8	9.7	408	266
	64	18.4	6.7	369	263
AMSE2	16	20.5	13.6	550	464
	32	15.3	7.6	340	210
	64	14.6	5.2	300	210

Table 2: Average mean squared errors ($\times 10^{-3}$) for hazard function estimation based on 200 repetitions of the simulations given in section 3.1 and 3.2 for sample sizes $n = 200$ and $n = 500$. The figures are based on the same data used to form table 1

		Hazard λ_1		Hazard λ_2	
	Bins K	$n = 200$	$n = 500$	$n = 200$	$n = 500$
AMSE	16	64.4	55.4	3050	3090
	32	78.6	55.4	4060	1820
	64	112.0	99.5	2080	1970
AMSE2	16	5.8	5.9	182	295
	32	2.6	2.1	152	66
	64	2.5	1.6	48	32

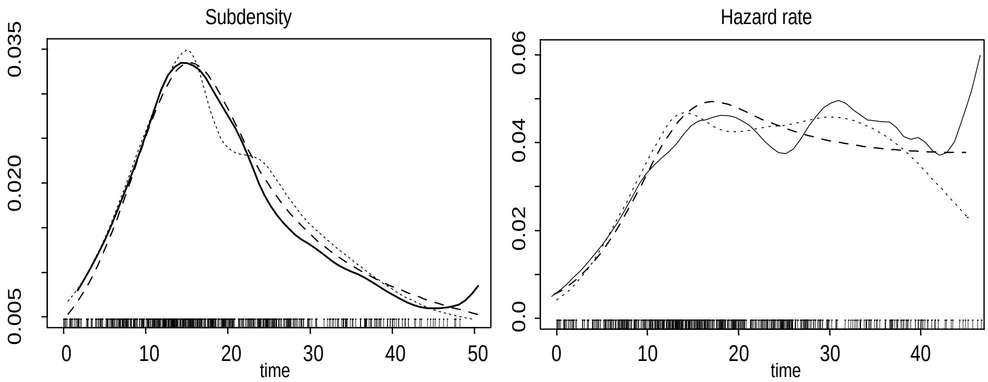

Figure 3: Wavelet (solid), Logspline (dashed) and Local polynomial (dotted) density and hazard rate estimates from the Liver Metastase data set. The bottom tick marks show the location of observed failures and the top ones show the location of censored observations.

the resolutions which minimize these pointwise MSEs vary considerably with where the point is located. Given this wide variation, the complexity of developing a true locally adaptive method for resolution selection, and given that cross-validation is not always reliable in individual samples, in exploratory analysis the best approach may be to examine several estimates using a relatively small range of resolutions.

The simulation results show that the number of intervals used has little effect on performance, at least for the fairly smooth models considered here. Similar results held for other configurations examined.

4 Examples

4.1 Liver metastases

As an example consider the data analyzed by Haupt and Mansmann (1995), giving survival times for patients with liver metastases from a colorectal primary tumour without other distant metastases. The survival times are measured in months. In this data set there are 363 uncensored and 259 censored observations. In order to compare our wavelet estimator (solid line) we also report on the figures the logspline subdensity estimator (dashed line) and the local polynomial estimator (dotted line) implemented in locfit (see Loader, 1995). Each estimation method uses a global data driven choice for the smoothing parameters (for pointers to an archive and for more details on this choice for logspline and locfit the reader is referred to the papers by Kooperberg and Stone, 1992 and Loader, 1995). The wavelet estimators (density and hazard rate) were computed as described in the previous sections, using 64 bins during the binning process and a resolution

$j(n) = 2$ chosen by folded cross validation. The logspline estimator was calculated by specifying that the density equals zero to the left of 0, using a BIC penalty and stepwise knot deletion. Four knots were selected for the spline that is fitted to the log-density. The local polynomial estimator was used with its default values and the fitted density had 1.167 equivalent degrees of freedom. Both logspline and locfit were applied on the un-binned data.

The left panel in Figure 3 gives estimated densities for observed failures, while the right one displays the corresponding hazard rate estimators.

The increasing risk at about 14 months is evident. On the hazard rates plot of figure 3 one can see that the estimates seem reliable on a smaller interval than the one that extends beyond 40 months. The wavelet estimate exhibits some boundary effects due to the fact that few only cases are still at risk after 40 months. Note also that the "constancy" in the hazard rate at later follow-up times observed in the logspline hazard rate estimate is supported by the wavelet estimator. The pronounced increase of the wavelet estimator is probably due to a boundary effect that is also noticeable to a lesser extent by the locfit estimator.

4.2 Employment example

Our second example is concerned with the unemployment dynamics of a population of 632 women. This data set has also been studied by Bonnal and Fougère (1990) using parametric methods. For a given individual the survival time is the time from when the individual is first unemployed until that person obtains employment. If an individual is still unemployed at the end of the study then that person is right censored (so time to employment is after the end of the study). In this data set there are 272 censored observations. We estimated the subdensities and hazard rates using our wavelet method, locfit and logspline. The wavelet estimators (density and hazard rate) were computed using 64 bins during the binning process and a resolution $j(n) = 3$ chosen by folded cross validation. The locfit and logspline procedures were used with their default settings, resulting in 2.01 fitted degrees of freedom and 5 knots, respectively.

Figure 4 reproduces the results. In Figure 4 it can be seen that the automatic logspline estimator, being based on only 5 knots, is oversmoothed. This is probably due to the placement and the selected number of knots.

One can reach very different conclusions by using each of the different estimators. According to the logspline hazard rate estimator, the instantaneous rate for transition from unemployment to work increases during a first period of about 3 months (90 days) and then decreases monotonically. For the locfit estimator after this first period, the rate seems to wander around a constant level, i.e. a women unemployed for a long period

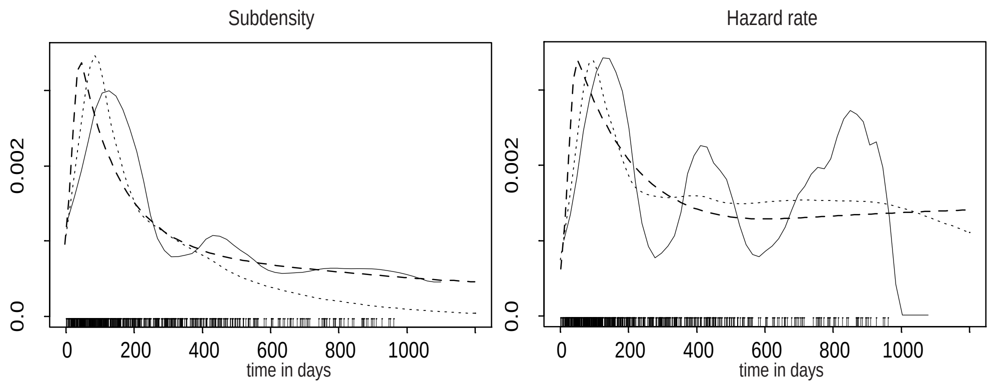

Figure 4: Wavelet (solid), Logspline (dashed) and Local polynomial (dotted) subdensities and hazard rate estimates for the unemployment data. The bottom tick marks show the location of observed failures and the top ones show the location of censored observations.

has the same chance as a recently unemployed women to find work. The results of the wavelet estimator are interpretable and interesting. The wavelet estimator indicates an increase in the hazard rate during the first three and half months with a strong decrease after (i.e. women are much more likely to find work in this period), reaching a fluctuating level after 360 days. During the months that follow one can discern some particular periods with a local increase of the hazard rate (the modes in the locfit hazard estimator are also apparent but less so). These increases may be related to other factors such as the end of indemnities for unemployment for proportions of the population.

In light of the example and the simulations and much additional experience with our density and hazard rate estimation method, we are convinced that the current implementation is of considerable practical value in data analysis. It is sufficiently accurate and flexible to handle peaks in the middle of the data (see Figure 2). However, it does not work very well far out in the tails (see Figure 3). A moderate amount of censoring has very little effect on the accuracy of estimation and the procedure can deal effectively with a high proportion of censoring. Moreover, the estimation procedure is rather insensitive to the choice of the initial binning width. Further, the comparison among the three methods of estimation is not totally fair since the wavelet resolution level is cross-validated while the others just use default smoothing parameters.

Conclusion

This paper introduced a wavelet-based method to estimate the density and hazard rate functions for right censored survival time data. We showed that our estimators possess pointwise and global mean-square consistency. Moreover we demonstrated that the estimators asymptotically obtain the best possible convergence rate in terms of MISE and also that they are asymptotically normally distributed.

Through simulations we demonstrate that the estimators perform well and provide some evidence to show that they are fairly insensitive to the number of bins. Two examples are presented that compare our estimator to established estimators. The comparisons reveal that the wavelet estimators compete favourably with the existing methods and their added flexibility provided new and useful interpretations of survival data.

Acknowledgments

We would like to thank the two referees and an associate editor for their very constructive comments and suggestions on earlier versions of this paper. Thanks are also due to Dr. Fanis Sapatinas who pointed out some last minute typographical errors. This research was supported by the CNRS-INRIA project IDOPT, Programme 6, at the Laboratoire IMAG-LMC of the University Joseph Fourier, Grenoble. Nason was supported in part by the U.K. EPSRC Grant GR/K70236. Nason would like to thank Laboratoire IMAG-LMC for support and excellent hospitality whilst visiting Grenoble to carry out this work.

Appendix

Proof of Lemma 3

Given the expectation of the U_k given by Lemma 1 we have:

$$\begin{aligned} \mathbb{E}[\tilde{f}_N^*(t)] &= 2^{-N/2} \sum_{k=0}^{K} \mathbb{E}\left[\frac{U_k}{\Delta}\right] \phi_{N,k}(t) = 2^{-N/2} \sum_{k=0}^{K} \left[f^*(t_k) + \mathcal{O}(\Delta)\right] \phi_{N,k}(t) \\ &= 2^{-N/2} \sum_{k=0}^{K} f^*(t_k)\phi_{N,k}(t) + \mathcal{O}(\Delta) 2^{-N/2} \sum_{k=0}^{K} \phi_{N,k}(t). \end{aligned}$$

Using the fact that $\sup_t \sum_k |\phi(t-k)| = M$, where M is a constant depending only on the scaling function ϕ, and since $\phi_{N,k}(t) = 2^{N/2}\phi(2^N t - k)$, it is easy to see that $\sup_t 2^{-N/2} \sum_k |\phi_{N,k}(t)| = M$. Therefore

$$\begin{aligned} I\!\!E[\tilde{f}_N^*(t)] &= 2^{-N/2} \sum_{k=0}^{K} f^*(t_k)\phi_{N,k}(t) + \mathcal{O}(\Delta) \\ &= \sum_{k=0}^{K} \{< f^*, \phi_{N,k} > + \Delta^{1/2}\mathcal{O}(\Delta^m)\}\phi_{N,k}(t) + \mathcal{O}(\Delta) \quad \text{by Lemma 2} \\ &= \sum_{k=0}^{K} < f^*, \phi_{N,k} > \phi_{N,k}(t) + \mathcal{O}(\Delta^{m+1/2}) + \mathcal{O}(\Delta), \end{aligned}$$

where m is the order of differentiability of f^*.

To prove expression (11) note that

$$\text{Var}[\tilde{f}_N^*(t)] = 2^{-N} \sum_{k=0}^{K} \text{Var}\left(\frac{U_k}{\Delta}\right) \phi_{N,k}^2(t) + 2^{-N} \sum_{k=0}^{K} \sum_{\ell=0,\ell\neq k}^{K} \text{Cov}\left(\frac{U_k}{\Delta}, \frac{U_\ell}{\Delta}\right) \phi_{N,k}(t)\phi_{N,\ell}(t)$$

By Lemma 1 we have

$$\begin{aligned} \text{Var}[\tilde{f}_N^*(t)] &= 2^{-N} \sum_{k=0}^{K} \left[\frac{f^*(t_k)}{n\Delta} + \mathcal{O}(n^{-1})\right] \phi_{N,k}^2(t) \\ &\quad + 2^{-N} \sum_{k=0}^{K} \sum_{\ell=0,\ell\neq k}^{K} \left[-\frac{1}{n} f^*(t_k) f^*(t_\ell) + \mathcal{O}(\Delta n^{-1})\right] \phi_{N,k}(t)\phi_{N,\ell}(t) \end{aligned}$$

Since f^* is uniformly bounded on $[0, \tau]$ and since $\Delta = \mathcal{O}(2^{-N})$ it follows that:

$$\begin{aligned} \text{Var}[\tilde{f}_N^*(t)] &\leq \frac{1}{n} C_1 \sum_{k=0}^{K} \phi_{N,k}^2(t) + \mathcal{O}(\Delta/n) \sum_{k=0}^{K} \phi_{N,k}^2(t) \\ &\quad + C_2 \frac{\Delta}{n} \sum_{k} \sum_{k\neq\ell} |\phi_{N,k}(t)\phi_{N,\ell}(t)| + \mathcal{O}(\Delta^2/n) \sum_{k} \sum_{k\neq\ell} |\phi_{N,k}(t)\phi_{N,\ell}(t)|. \end{aligned}$$

Using the fact that ϕ is at least m-regular in the sense of Meyer (1990), we have

$$\sum_{k=0}^{K} \phi_{N,k}^2(t) = \mathcal{O}(2^N) \quad \text{and} \quad \sum_{k} \sum_{k\neq\ell} |\phi_{N,k}(t)|\,|\phi_{N,\ell}(t)| = \mathcal{O}(2^{2N}).$$

where $\mathcal{O}$ does not depend on t. It then follows that

$$\text{Var}[\tilde{f}_N^*(t)] = \mathcal{O}\left\{(n\Delta)^{-1}\right\} + \mathcal{O}(n^{-1}),$$

which completes the proof. ∎

Proof of Theorem 1

Write

$$\begin{aligned} E(t) &= \hat{f}_n(t) - f^*(t) = \Pi_{j(n)}(\tilde{f}_N^*)(t) - f^*(t) \\ &= \Pi_{j(n)}(\tilde{f}_N^*)(t) - \Pi_{j(n)}(f^*)(t) + \Pi_{j(n)}(f^*)(t) - f^*(t) = S(t) + A(t) \end{aligned}$$

with S denoting the stochastic error and A the approximation error of the estimate. Using the above expression, the fact that $m \geq 1$ and the results of Lemma 3 we have:

$$\mathbb{E}[E(t)] = \mathbb{E}[S(t)] + A(t) = \Pi_{j(n)}(\Pi_N(f^*))(t) + \mathcal{O}(\Delta) - \Pi_{j(n)}(f^*)(t) + A(t).$$

Now, since $V_{j(n)} \subset V_N$, it follows that

$$\mathbb{E}[E(t)] = \mathcal{O}(\Delta) + A(t).$$

Using the fact $\|\Pi_{j(n)}(f^*) - f^*\|_2 = o(2^{-mj(n)})$ (see (8)), the squared integrated bias of $\hat{f}_n$ behaves like

$$\mathcal{O}(\Delta^2) + \mathcal{O}(2^{-2mj(n)}).$$

As for the variance part of the MISE, say V, observe first that

$$\begin{aligned} \Pi_{j(n)}(\tilde{f}_N^*)(t) &= \sum_{k=0}^{2^{j(n)}-1} < \tilde{f}_N^*, \phi_{j(n),k} > \phi_{j(n),k}(t) \\ &= \sum_{k=0}^{2^{j(n)}-1} \left\{ \sum_{\ell=0}^{2^N-1} 2^{-N/2} < \frac{U_\ell}{\Delta}\phi_{N,\ell}, \phi_{j(n),k} > \right\} \phi_{j(n),k}(t) \end{aligned}$$

It therefore follows that

$$\Pi_{j(n)}(\tilde{f}_N^*)(t) - \Pi_{j(n)}(\mathbb{E}(\tilde{f}_N^*(t))) = \sum_k 2^{-N/2} \sum_\ell \frac{U_\ell - \mathbb{E}(U_\ell)}{\Delta} < \phi_{N,\ell}, \phi_{j(n),k} > \phi_{j(n),k}(t).$$

Using Parseval's relation we have:

$$\int \left\{ \Pi_{j(n)}(\tilde{f}_N^*)(t) - \Pi_{j(n)}\mathbb{E}(\tilde{f}_N^*)(t)) \right\}^2 dt = 2^{-N} \sum_\ell \left\{ \sum_k \frac{U_k - \mathbb{E}(U_k)}{\Delta} < \phi_{N,k}, \phi_{j(n),\ell} > \right\}^2. \tag{14}$$

Taking expectations in both sides of (14) and using the results of Lemma 1 yields

$$\begin{aligned} V &= \sum_\ell 2^{-N} \sum_k \mathbb{E}\left\{ \frac{(U_k - \mathbb{E}(U_k))^2}{\Delta^2} \right\} < \phi_{N,k}, \phi_{j(n),\ell} >^2 \\ &+ \sum_\ell 2^{-N} \sum_k \sum_{h \neq k} \text{Cov}\{\frac{U_k}{\Delta}, \frac{U_h}{\Delta}\} < \phi_{N,k}, \phi_{j(n),\ell} >< \phi_{N,h}, \phi_{j(n),\ell} > \end{aligned} \tag{15}$$

By Lemma 1 the first part in the right hand side of (15) is equal to

$$\Delta \sum_{k,\ell} \{ \frac{f^*(t_k)}{n\Delta} + \mathcal{O}(n^{-1}) \} < \phi_{N,k}, \phi_{j(n),\ell} >^2 \tag{16}$$

Now, observe that

$$\sum_k < \phi_{N,k}, \phi_{j(n),\ell} >^2 = \|\Pi_N(\phi_{j(n),\ell})\|_2^2 = \|\phi_{j(n),\ell}\|_2^2 = 1. \tag{17}$$

Therefore, using again the fact that f^* is uniformly bounded, we see that (16) behaves like:

$$\mathcal{O}\left(\frac{2^{j(n)}}{n}\right) + \mathcal{O}(\Delta 2^{j(n)} n^{-1}) = \mathcal{O}\left(\frac{2^{j(n)}}{n}\right). \tag{18}$$

The second term in (15) is equal to

$$\sum_\ell 2^{-N} \sum_k \sum_{h \neq k} \{ -\frac{1}{n} f^*(t_k) f^*(t_h) + \mathcal{O}(\Delta n^{-1}) \} < \phi_{N,k}, \phi_{j(n),\ell} >< \phi_{N,h}, \phi_{j(n),\ell} > .$$

Now, by using (17)

$$\begin{aligned} \sum_k \sum_{h \neq k} < \phi_{N,k}, \phi_{j(n),\ell} >< \phi_{N,h}, \phi_{j(n),\ell} > &\leq \sum_k \sum_h | < \phi_{N,k}, \phi_{j(n),\ell} > | \, | < \phi_{N,h}, \phi_{j(n),\ell} > | \\ &= \left(\sum_k | < \phi_{N,k}, \phi_{j(n),\ell} > | \right)^2 \\ &\leq 2^N \sum_k < \phi_{N,k}, \phi_{j(n),\ell} >^2 = 2^N \|\phi_{j(n),\ell}\|_2^2 = 2^N \end{aligned}$$

Finally, it is easy to see that (16) behaves like

$$\mathcal{O}(\frac{2^{j(n)}}{n}) + \mathcal{O}(\frac{\Delta 2^{j(n)}}{n}) = \mathcal{O}(\frac{2^{j(n)}}{n}),$$

and this, together with (17), completes the proof. ∎

Proof of Theorem 2

First write

$$\sqrt{n\Delta}(\hat{f}_n(t) - f^*(t)) = \sqrt{n\Delta}(\hat{f}_n(t) - \mathbb{E}[\hat{f}_n(t)]) + \sqrt{n\Delta}(\mathbb{E}[\hat{f}_n(t)] - f^*(t)).$$

By the proof of Theorem 1 it is known that

$$\mathbb{E}[\hat{f}_n(t)] - f^*(t) = \mathcal{O}(\Delta) + (\Pi_{j(n)}(f^*(t)) - f^*(t)).$$

From Walter (1992), Theorem 4.1, p. 937, it follows that

$$\sup_{f\in H_m} \|\Pi_{j(n)}(f) - f\|_\infty = \mathcal{O}(2^{-j(n)(m-1/2)}).$$

It follows that $(\Pi_{j(n)}(f^*)(t) - f^*(t)) = \mathcal{O}(2^{-j(n)(m-1/2)})$ and consequently

$$\mathbb{E}[\hat{f}_n(t)] - f^*(t) = \mathcal{O}(\Delta) + \mathcal{O}(2^{-j(n)(m-1/2)}).$$

Consider now the following expansion

$$\begin{aligned} & \sqrt{n\Delta}(\hat{f}_n(t) - \mathbb{E}[\hat{f}_n(t)]) \\ = \; & \sqrt{n\Delta}(\hat{f}_n(t) - \tilde{f}_N^*(t)) + \sqrt{n\Delta}(\tilde{f}_N^*(t) - \mathbb{E}\tilde{f}_N^*(t)) + \sqrt{n\Delta}(\mathbb{E}\tilde{f}_N^*(t) - \mathbb{E}\hat{f}_n(t)) \\ = \; & I + II + III \quad \text{say.} \end{aligned}$$

The key argument in our proof is that II is asymptotically Gaussian while the terms I and III are vanishing as n goes to infinity.

Let us first show that I and III tend to zero. This basically follows from Walter's result quoted above. For the term I we have

$$\hat{f}_n(t) - \tilde{f}_N^*(t) = \Pi_{j(n)}(\tilde{f}_N^*)(t) - \tilde{f}_N^*(t) = \mathcal{O}(2^{-j(n)(m-1/2)}),$$

and the convergence follows from the assumptions of the theorem. For the term III we observe that

$$\left|\mathbb{E}\tilde{f}_N^*(t) - \mathbb{E}\hat{f}_n(t)\right| \le \mathbb{E}\left|\tilde{f}_N^*(t) - \hat{f}_n(t)\right| = \mathcal{O}(2^{-j(n)(m-1/2)}),$$

where we have used the fact that Walter's result implies that $|\Pi_{j(n)}f - f| \le K2^{-j(n)(m-1/2)}$ where K does not depend on f.

Finally, it remains to show that, given our assumptions, $\tilde{f}_N^*(t)$ is asymptotically Gaussian. The proof will rely on the decomposition

$$\sqrt{n\Delta}(\tilde{f}_N^*(t) - \mathbb{E}\tilde{f}_N^*(t)) = \sum_{i=1}^n \frac{1}{\sqrt{n\Delta}} \sum_{k=0}^K (Y_{ik} - p_k)\phi(2^N t - k) = \sum_{i=1}^n Z_{ni},$$

where $p_k = \mathbb{E}(Y_{ik})$.

It is important to note that, since ϕ is compactly supported, the sum over k only involves a finite number (depending on t) of terms. From arguments similar to those used in the proof of Lemma 3, we obtain

$$\mathrm{Var}[Z_{ni}] = \frac{1}{n}\left(\sum_{k=0}^K f^*(t_k)\phi^2(2^N t - k) + \mathcal{O}(\Delta)\right).$$

By continuity of f and since t is dyadic, it follows that

$$\mathrm{Var}[\sqrt{n\Delta}(\tilde{f}_N^*(t) - \mathbb{E}\tilde{f}_N^*(t))] \to f^*(t) \sum_{k=-\infty}^{\infty} \phi^2(k).$$

Note that when t is non dyadic the sum over k is not asymptotically stable (see Antoniadis *et al.*, (1994) for a similar remark when estimating a regression function by wavelet methods). To obtain asymptotic normality it remains to check the Lindeberg condition. Setting $U_{ni} = Z_{ni}/(\text{Var}Z_{ni})^{1/2}$, this amounts in showing that

$$\mathbb{E}\left\{U_{ni}^2 \mathbf{I}_{|U_{ni}|>\varepsilon\sqrt{n}}\right\} \to 0.$$

By the Cauchy-Schwarz and Chebyshev inequalities, we have

$$\mathbb{E}\left\{U_{ni}^2 \mathbf{I}_{|U_{ni}|>\varepsilon\sqrt{n}}\right\} \leq \left\{\mathbb{E}(U_{ni})^4\right\}^{1/2} (\varepsilon\sqrt{n})^{-1}.$$

Straightforward calculations yield

$$\mathbb{E}(Z_{ni}^4) = \mathcal{O}\left(\frac{1}{n^2\Delta}\right),$$

and $\mathbb{E}(U_{ni}^4) = \mathcal{O}(n^2)\mathbb{E}(Z_{ni}^4)$. Therefore,

$$\mathbb{E}\left\{U_{ni}^2 \mathbf{I}_{|U_{ni}|>\varepsilon\sqrt{n}}\right\} = \mathcal{O}\left(\frac{1}{\sqrt{n\Delta}}\right),$$

and this completes the proof. ∎

Proof of Proposition 2

Let us first note that

$$\begin{aligned}\hat{\lambda}_n(t) &= \left[\mathbb{E}(\hat{f}_n(t)) + \{\hat{f}_n(t) - \mathbb{E}(\hat{f}_n(t))\}\right] \{\mathbb{E}(1-\hat{L}_n(t))\}^{-1} \\ &\quad \times \left\{1 + [\mathbb{E}(\hat{L}_n(t)) - \hat{L}_n(t)](\mathbb{E}(1-\hat{L}_n(t)))^{-1}\right\}^{-1}.\end{aligned}$$

By the Chung-Smirnov property of $\hat{L}_n$ and using Taylor's theorem it follows that

$$\begin{aligned}\hat{\lambda}_n(t) &= \mathbb{E}(\hat{f}_n(t))\{\mathbb{E}(1-\hat{L}_n(t))\}^{-1} + \{\hat{f}_n(t) - \mathbb{E}(\hat{f}_n(t))\}\{\mathbb{E}(1-\hat{L}_n(t))\}^{-1} \\ &\quad - \left\{\mathbb{E}(\hat{L}_n(t)) - \hat{L}_n(t)\right\} \mathbb{E}(\hat{f}_n(t)) \left(\mathbb{E}(1-\hat{L}_n(t))\right)^{-2} \\ &\quad + o\left[|\hat{f}_n(t) - \mathbb{E}(\hat{f}_n(t))| + |\mathbb{E}(\hat{L}_n(t)) - \hat{L}_n(t)|\right].\end{aligned}$$

Now, using again Taylor's formula, we have

$$\begin{aligned}\frac{\mathbb{E}(\hat{f}_n(t))}{\mathbb{E}(1-\hat{L}_n(t))^{-1}} &= \left(f^*(t) + [\mathbb{E}(\hat{f}_n(t)) - f^*(t)]\right)(1-L(t))^{-1} \\ &\quad \times \left\{1 - \frac{\mathbb{E}(\hat{L}_n(t)) - L(t)}{1-L(t)} + o\left(|\mathbb{E}(\hat{L}_n(t)) - L(t)|\right)\right\}.\end{aligned}$$

From the above expressions and the fact that $\lambda(t) = f^*(t)/(1-L(t))$, routine calculations show that the mean squared error (and the MISE as well) of $\hat{\lambda}_n(t)$ is the same as the one of

$$\left[\frac{\hat{f}_n(t) - f^*(t)}{1 - L(t)} - \frac{f^*(t)}{(1 - L(t))^2}[L(t) - \hat{L}_n(t)]\right]. \qquad (19)$$

The assertion of the proposition is now easily seen to be a consequence of Theorem 1 and Proposition 1 and the fact that $(1 - L(t))$ is uniformly bounded away from 0 on $[0, \tau]$ by definition of τ. ∎

References

Andersson, L., Hall, N., Jawerth, B. and Peters, G. (1993) Wavelets on closed subsets of the real line. In *Topics in the theory and applications of wavelets* (L.L. Schumacher and G. Webb eds.) London: Academic Press.

Antoniadis, A. (1989) A penalty method for nonparametric estimation of the intensity of a counting process. *Ann. Inst. Math. Statist.*, **41**, 781–807.

Antoniadis, A. and Grégoire, G. (1990) Penalized likelihood estimation for rates with censored survival data. *Scand. J. Statist.*, **17**, 43–63.

Antoniadis, A., Grégoire, G. and McKeague, I. (1994) Wavelet methods for curve estimation. *J. Am. Statist. Ass.*, **89**, 1340–1353.

Antoniadis, A. (1994) Smoothing noisy data with coiflets. *Statistica Sinica*, **4**, 651–678.

Antoniadis, A. (1996) Smoothing noisy data with tapered coiflet series. *Scand. J. Statist.*, **23**, 313–330.

Beran, R. (1981) Nonparametric regression with randomly censored survival data. *Technical Report.* Department of Statistics, University of California, Berkeley.

Bonnal, L. and Fougère, D. (1990) Les déterminants individuels de la durée de chomâge. *Economie et Prévision*, **96**, 113–129.

Buckheit, J.B. and Donoho, D.L. (1995) Wavelab and reproducible research. In *Wavelets and statistics*, Lecture Notes in Statistics, 103 (A. Antoniadis and G. Oppenheim, eds.) Springer-Verlag.

Carbonez, A., Györfi, L. and Van der Meulen, E.C. (1995) Partitioning estimates of a regression function under random censoring. *Statistics and Decisions*, **13**, 21–37.

Cohen, A., Daubechies, I. and Vial, P. (1993) Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harm. Anal.*, **1**, 54–82.

Dabrowska, D.M. (1987) Non-parametric regression with censored survival time data. *Scand. J. Statist.*, **14**, 181–197.

Daubechies, I. (1992) *Ten lectures on wavelets.* Philadelphia: SIAM.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D.L. and Johnstone, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

Donoho, D.L, Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia (with discussion)? *J.R. Statist. Soc.* B, **57**, 301–369.

Dvoretsky, A., Kiefer, J. and Wolfowitz, J. (1956) Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, **23**, 277–281.

Efromovitch, S.Y. (1989) Sequential nonparametric estimation of a density. *Theor. Prob. Appl.*, **34**, 228–239.

Gray, R.J. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Am. Statist. Ass.*, **87**, 942–951.

Haupt, G. and Mansmann, U. (1995) Survcart: S and C code for CART analysis with survival data. A `statlib` shar archive.

Härdle, W. (1991) *Smoothing techniques with implementation in S.* New York: Springer-Verlag.

Härdle, W. and Scott, D.W. (1992) Smoothing by weighted averaging of rounded points. *Comput. Statist.*, **7**, 97–128.

Ibragimov, I.A. and Khasminski, R.Z. (1982) Bounds for the risk of nonparametric regression estimates. *Theor. Prob. Appl.*, **27**, 84–99.

Kronmal, R. and Tarter, M. (1968) The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Stat. Assoc.*, **63**, 925–952.

Kooperberg, C. and Stone, C.J. (1992) Logspline density estimation for censored data. *J. Comput. Graph. Statist.*, **1**, 301–328.

Kronmal, R. and Tarter, M. (1968) The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Stat. Assoc.*, **63**, 925–952.

Loader, C. (1995) Locfit: A program for local fitting. *Technical report.* AT&T Bell Laboratories. Available via anonymous ftp at the URL: `http://cm.bell-labs.com/stat/project/locfit/index.html`

Mallat, S. (1989) Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Am. Math. Soc.*, **315**, 69–87.

Nason, G.P. (1996) Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc.*, B, **58**, 463–479.

Nason, G.P. and Silverman, B.W. (1994) The discrete wavelet transform in S. *J. Comput. Graph. Statist.*, **3**, 163–191.

Nussbaum, M. (1985) Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.*, **13**, 984–997.

Puri, M.L. and Ralescu, S.S. (1986) Central limit theorem for perturbed empirical distribution functions evaluated at a random point, *J. Multivariate Anal.*, **19**, 273–279.

O'Sullivan, F. (1988) Fast computation of fully automated log-density and log-hazard estimators, *SIAM J. Sci. and Statist. Comp.*, **9**, 363–379.

Stone, C.J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040–1053.

Tanner, M.A. and Wong, W.H. (1984) Data-based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis. *J. Amer. Stat. Assoc.*, **79**, 174–182.

Tribouley, K. (1995) Adaptive density estimation. In *Wavelets and Statistics*, Eds. A. Antoniadis and G. Oppenheim, Lecture Notes in Statistics, **103**, 385–396, Springer-Verlag, New York.

Walter, G.G. (1992) Approximation of the delta function by wavelets, *J. Approx. Theor.*, **71**, 329–343.

Winter, B.B. (1973) Strong uniform consistency of integrals of density estimators, *Can. J. Statist.*, **1**, 247–253.

Winter, B.B. (1979) Convergence rate of perturbed empirical distribution functions, *J. Appl. Prob.*, **16**, 163–173.

Yukish, J.E. (1989) A note on limit theorems for perturbed empirical processes. *Stoch. Proc. Appl.*, **33**, 163–173.