

On wavelet estimation in censored regression

Linyuan Li

Department of Mathematics and Statistics, University of New Hampshire, USA

Brenda MacGibbon

Département de Mathématiques, Université du Québec à Montréal, CANADA

and

Christopher Valenta

Information Resources Inc, New Jersey, USA

Abstract

The Cox proportional hazards model has become the model of choice to use in analyzing the effects of covariates on survival data. However, this assumption has significant restrictions on the behavior of the conditional survival function. The accelerated failure time model, which models the survival time and covariates directly through regression, provides an alternative approach to interpret the relationship between survival times and covariates. We consider here the estimation of the nonparametric regression function in the accelerated failure time model under right random censorship and investigate the asymptotic rates of convergence of estimators based on thresholding of empirical wavelet coefficients. We show that the estimators achieve nearly optimal minimax convergence rates within logarithmic terms over a large range of Besov function classes B_{pq}^α , $\alpha > 1/p$, $p \geq 1$, $q \geq 1$, a feature not available for the linear estimators when $p < 2$. The performance of the estimators is tested via simulation and the method is applied to the Stanford Heart Transplant data.

Short title: Wavelets in censored regression

2000 Mathematics Subject Classification: Primary: 62G07; Secondary: 62G20

Keywords: Adaptive estimation; censored data; minimax estimation; nonlinear wavelet-based estimator; nonparametric regression; rates of convergence

1 Introduction

The Cox proportional hazards model has become the most popular approach to analyze the effects of covariates on survival data. However, this assumption has significant restrictions on the behavior of the conditional survival function (for instance, see Portnoy, 2003 and the references cited therein). The accelerated failure time model, which models the survival time (or a transformation of the survival time) and covariates directly through regression rather than modelling the conditional survival and hazard functions, provides a valuable complement to interpreting the relationship between survival time and covariates. Formally, let Y be a random variable representing the survival time of a subject taking part in a medical or other experimental study and X be a random variable of covariate, *e.g.*, age, sex, blood pressure, *etc.* In regression analysis we want to estimate Y given X , *i.e.*, to estimate the mean regression function $g(x) = E(Y|X = x)$, which is equivalent to estimating the function g from the regression model: $Y = g(X) + \sigma(X)\varepsilon$, where $\sigma(\cdot)$ is the conditional variance representing possible heteroscedasticity and ε represents the random error, which is assumed to be *i.i.d.*

In industrial life-testing, medical research and other studies, the observation of the occurrence of a failure may be made impossible by the previous occurrence of a censoring event, such as the termination of the study or withdrawal from the study. In this case only some of the observations represent true failure times. More precisely, let Y_1, Y_2, \dots, Y_n denote the survival times and X_1, X_2, \dots, X_n the associated covariates, and let us assume that $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$ are independent and have a joint distribution function (*d.f.*) $F^0(x, y)$. Also let T_1, T_2, \dots, T_n denote the *i.i.d.* censoring times with a common *d.f.* G . It is assumed that (Y_i, X_i) is independent of T_i for each i . Rather than observing $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$, the variables of interest, in the randomly right-censored model, (Z_i, δ_i, X_i) is observed, where $Z_i = \min(Y_i, T_i) = Y_i \wedge T_i$ and $\delta_i = I(Y_i \leq T_i)$, $i = 1, 2, \dots, n$, where $I(A)$ denotes the indicator function of the set A . We denote H as the *d.f.* of Z_1 , and $\tau_H = \inf\{x : H(x) = 1\} \leq \infty$ is the least upper bound for the support of H .

Many authors have assumed that the mean of the log-lifetime is a linear function of the covariate and have estimated the linear regression parameters α and β (see, *e.g.*, Miller, (1976); Buckley and James (1979); Koul *et al.* (1981); Miller *et al.* (1982)). Stute (1999) extended the above linear model to a nonlinear one with parameter θ and proved

the consistency and asymptotic normality of the weighted least square estimator of θ .

Here, we consider the nonparametric regression problem :

$$Y_i = g(X_i) + \sigma(X_i)\varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1.1)$$

where $\varepsilon_1, \dots, \varepsilon_n$ are *i.i.d.* errors with mean 0. Our goal is to estimate the unknown function g based on the observations $\{(X_i, Z_i, \delta_i) : i = 1, 2, \dots, n\}$, a random sample from the population (X, Z, δ) .

Although there is a vast literature in nonparametric regression, (see, *e.g.*, the books of Härdle (1990), Fan and Gijbels (1996) and the references cited therein), these methods are not directly applicable to censored data. Here we use ideas from Buckley and James (1979), Koul *et al.* (1981) and Fan and Gijbels (1994) to transform the censored data in an unbiased way. Once such a transformation is carried out, the regression function can be estimated as if the complete data were observed. Fan and Gijbels (1994) considered a local linear regression smoother and derive its asymptotic normality for a fixed smooth regression function.

In many examples of survival data in the medical literature, however, the regression function may exhibit spatially inhomogeneous smoothness or discontinuities. For completely observed data, it is well known that wavelet estimators can handle discontinuities, have extraordinary local adaptability and typically achieve optimal convergence rates over exceptionally large function spaces. (We cite Donoho *et al.* (1994, 1995, 1996), Donoho and Johnstone (1998), Hall and Patil (1995, 1996a, b) and Hall *et al.* (1998, 1999a, b)) on the performance of nonlinear wavelet estimators. In all of the above research it is assumed that the observations are complete. For the censored nonparametric regression model here, we adapt the wavelet estimators to right randomly censored data by first transforming the data as in Fan and Gijbels (1994). We assume that the regression function g belongs to a large function class of Besov spaces B_{pq}^α and then show that the wavelet estimators achieve nearly optimal minimax convergence rates within logarithmic terms over B_{pq}^α .

In the next section, we give the necessary definitions and define the nonlinear wavelet-based mean regression function estimators. The main results with proofs are described in Section 3. Section 4 contains a modest simulation study and Section 5 illustrates the wavelet estimator using the Stanford Heart Transplant data. Section 6 is the conclusion. The proofs of some technical results are given in the appendix.

2 Notation and Estimators

As is usual in the wavelet literature we assume that the regression function $g(x)$ is defined for $x \in (0, 1)$. Our aim is to estimate g , by non-linear thresholding of the empirical wavelet coefficients. We assume that the father and mother wavelets, $\phi(x)$ and $\psi(x)$, are bounded and compactly supported, and $\int \phi = 1$. We call a wavelet ψ *r-regular* if ψ has r vanishing moments and r continuous derivatives. Let

$$\phi_{j_0 k}(x) = 2^{j_0/2} \phi(2^{j_0} x - k), \quad \psi_{jk}(x) = 2^{j/2} \psi(2^j x - k), \quad x \in \mathbb{R}, \quad j_0, j \in \mathbb{Z}.$$

Then the collection $\{\phi_{j_0 k}, \psi_{jk}, j \geq j_0, k \in \mathbb{Z}\}$ is an orthonormal basis (ONB) of $L^2(\mathcal{R})$. (For the existence and properties of such wavelets, we cite Daubechies (1992) and Cohen *et al.* (1993)). Therefore, for all $f \in L^2(\mathcal{R})$,

$$f(x) = \sum_{k \in \mathcal{Z}} \alpha_{j_0 k} \phi_{j_0 k}(x) + \sum_{j \geq j_0} \sum_{k \in \mathcal{Z}} \beta_{jk} \psi_{jk}(x),$$

where

$$\alpha_{j_0 k} = \int f(x) \phi_{j_0 k}(x) dx, \quad \beta_{jk} = \int f(x) \psi_{jk}(x) dx.$$

Our goal is to study wavelet-based estimators' asymptotic rates of convergence over a large range of Besov function classes B_{pq}^α , $\alpha > 0$, $1 \leq p, q \leq \infty$. Besov spaces form a very rich class of function spaces, (which include, in particular, the well-known Sobolev and Hölder spaces of smooth functions H^m and C^s (B_{22}^m and $B_{\infty, \infty}^s$ respectively), as well as function classes of significant spatial inhomogeneity such as the Bump Algebra and Bounded Variations Classes of Triebel (1992). We consider here the intersection of $B_\infty(A) = \{g : \|g\|_\infty \leq A\}$ with the following subset of the Besov space B_{pq}^α (where $\alpha p > 1$, $p, q \in [1, \infty]$) :

$$\mathcal{G}_{pq}^\alpha(M) = \{g : g \in B_{pq}^\alpha, \|g\|_{B_{pq}^\alpha} \leq M, \text{supp } g \subseteq [0, 1]\} ;$$

i. e., $\mathcal{G}_{pq}^\alpha(M)$ is a subset of functions with fixed compact support and bounded in the norm of the Besov space B_{pq}^α . Moreover, $\alpha p > 1$ implies \mathcal{G}_{pq}^α is a subset of the space of bounded continuous functions. For a given *r-regular* mother wavelet ψ with $r > \alpha$, we use the sequence norm of the wavelet coefficients of a function $f \in B_{pq}^\alpha$ by

$$|f|_{B_{pq}^\alpha} = \left(\sum_k |\alpha_{j_0 k}|^p \right)^{1/p} + \left\{ \sum_{j=j_0}^{\infty} \left[2^{js} \left(\sum_k |\beta_{jk}|^p \right)^{1/p} \right]^q \right\}^{1/q}. \quad (2.1)$$

where $s = \alpha + \frac{1}{2} - \frac{1}{p}$, to evaluate $\|f\|_{B_{p,q}^\alpha}$, since Meyer (1992) has shown their equivalence.

For such a mother wavelet ψ , the wavelet expansion of $g \in \mathcal{G}_{pq}^\alpha$, is

$$g(x) = \sum_{k=0}^{2^{j_0}-1} \alpha_{j_0k} \phi_{j_0k}(x) + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \beta_{jk} \psi_{jk}(x), \quad x \in [0, 1], \quad (2.2)$$

where

$$\alpha_{j_0k} = \int g(x) \phi_{j_0k}(x) dx, \quad \beta_{jk} = \int g(x) \psi_{jk}(x) dx.$$

In order to define our threshold estimator and to use the approximation of the Kaplan-Meier integral in Lemma 3.1, we must assume that the survival time Y is bounded by a positive constant B . We also assume $P(T > B) > 0$, *i.e.*, $G(B) < 1$. Under these assumptions, we have $\tau_F = \tau_H$, where τ_F is the least upper bound for the support of F . In most real life examples this boundedness assumption is valid.

The proposed nonlinear wavelet estimator of $g(x)$ is

$$\hat{g}(x) = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0k} \phi_{j_0k}(x) + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > d\sqrt{n^{-1} \ln n}) \psi_{jk}(x), \quad (2.3)$$

where the smoothing parameters j_0, j_1 satisfying $2^{j_0} \simeq \log_2 n$ and $2^{j_1} \simeq n(\log_2 n)^{-2}$, where the notation $2^{j(n)} \simeq h(n)$ indicates that $j(n)$ is chosen to satisfy the inequalities $2^{j(n)} \leq h(n) < 2^{j(n)+1}$. (For the sake of simplicity, we always omit the dependence of j_0 and j_1 on n). The threshold constant $d = \sqrt{2(1 - G(B))^{-1}} B$, where B is related to the boundness of Y and $\hat{\alpha}_{j_0k}$ and $\hat{\beta}_{jk}$ are defined as follows:

$$\hat{\alpha}_{j_0k} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i Z_i \phi_{j_0k}(X_i)}{1 - \hat{G}_n(Z_i-)}, \quad \hat{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i Z_i \psi_{jk}(X_i)}{1 - \hat{G}_n(Z_i-)}. \quad (2.4)$$

Here \hat{G}_n denote the Kaplan-Meier estimator of the *d.f.* G , *i.e.*,

$$\hat{G}_n(x) = 1 - \prod_{i=1}^n \left[1 - \frac{1 - \delta_{(i)}}{n - i + 1} \right]^{I(Z_{(i)} \leq x)},$$

where, $Z_{(i)}$ is the i -th ordered Z -value and $\delta_{(i)}$ is the concomitant of the i -th order Z statistic, *i.e.*, $\delta_{(i)} = \delta_j$ if $Z_{(i)} = Z_j$.

If there is no covariate, then the Kaplan-Meier estimator of the *d.f.* F based on (Z_i, δ_i) , $i = 1, 2, \dots, n$ is

$$\hat{F}_n(x) = 1 - \prod_{i=1}^n \left[1 - \frac{\delta_{(i)}}{n - i + 1} \right]^{I(Z_{(i)} \leq x)}.$$

Note that $\delta_i/n(1 - \hat{G}_n(Z_i-))$ is the jump of the Kaplan-Meier estimator \hat{F}_n at Z_i . In the presence of a covariate, Stute (1993) extends \hat{F}_n so as to obtain a consistent estimator \hat{F}_n^0 of the joint *d.f.* $F^0(x, y) = P(X \leq x, Y \leq y)$ of (X, Y) when Y is subject to censoring and X is observable, i.e.,

$$\hat{F}_n^0(x, y) = \sum_{i=1}^n \frac{\delta_i}{n(1 - \hat{G}_n(Z_i-))} I(X_{(i)} \leq x, Z_{(i)} \leq y), \quad (2.5)$$

where $X_{(i)}$ is the concomitant variable associated with the i -th order statistic Z_i . Hence we may write the empirical wavelet coefficients as $\hat{\beta}_{jk} = \int y \psi_{jk}(x) \hat{F}_n^0(dx, dy)$, which will be used in the proof of Lemma 3.1.

3 Main results

In the main theorem below, we study the estimator defined by equations (2.3) and (2.4) and establish its convergence rate over a large function class. Typically, one first constructs a wavelet-based estimator of $g_1(x) = \int y f(x, y) dy = f(x)g(x)$, where $f(x)$ and $f(x, y)$ are probability density functions of random variables X and (X, Y) respectively. Then the estimation of the mean regression function g can be obtained by dividing the estimator of f , which can be estimated in a variety of ways. For simplicity of exposition, we assume the design variable X has a uniform density over $(0, 1)$. Similar results, however, can be derived for the non-uniform stochastic design case. The following theorem shows that the wavelet-based estimator, based on simple thresholding of the empirical wavelet coefficients, attains a nearly optimal convergence rates over a large range of Besov function classes and is completely adaptive; *i.e.*, it behaves as if it knows in advance in which class the function lies.

Theorem 3.1. *Suppose the wavelet ψ is r -regular. Then, there exists a constant D , such that for all $M, A \in (0, \infty)$; $1/p < \alpha < r$; $p, q \in [1, \infty]$,*

$$\sup_{g \in \mathcal{G}_{pq}^\alpha(M) \cap B_\infty(A)} E \int (\hat{g} - g)^2 \leq D \left(\frac{\log_2 n}{n} \right)^{2\alpha/(1+2\alpha)},$$

where \hat{g} is defined by equations (2.3) and (2.4).

The method of proof for the above theorem is similar to that of Theorem 3.1 of Li (2004), which considers the density estimator with randomly censored data. The

difference is that here in the regression case, there exists a covariate yielding different estimators. The overall proof of the theorem follows along the lines of Donoho *et al.* (1996) and Hall *et al.* (1998) for the complete data case, while overcoming some technical difficulties encountered from censored data. For complete data, the empirical wavelet coefficients would be defined as $\hat{\beta}_{jk} = n^{-1} \sum_{i=1}^n Y_i \psi_{jk}(X_i)$, which is an average of n i.i.d. random variables. In this case, it is easy to investigate the large deviation behavior of empirical coefficients $\hat{\beta}_{jk}$. For the censored data case, the empirical wavelet coefficients are constructed through the Kaplan-Meier estimators of the distribution functions as in (2.4). Hence they are no longer sums of i.i.d. random variables. The key part of the proof is to approximate the empirical coefficients $\hat{\beta}_{jk}$ with an average of i.i.d. random variables with a sufficiently small rate. Stute (1995) approximates the Kaplan-Meier integrals as an average of i.i.d. random variables with a certain rate in probability. Nevertheless we are able to show that the above approximations hold in L^2 also, since mean integrated squared error considers L^2 error. In order to prove the theorem, we need the following lemma. This result allows us to deal with empirical coefficients in censored data as in the complete data case.

Lemma 3.1. *Let $\hat{\alpha}_{j_0k}$ and $\hat{\beta}_{jk}$ be defined as in equations (2.4). Also, let*

$$\begin{aligned} \varphi_{j_0k}(x, y) &= y \phi_{j_0k}(x), \quad k = 0, 1, 2, \dots, 2^{j_0} - 1, \\ \varphi_{jk}(x, y) &= y \psi_{jk}(x), \quad j = j_0, j_0 + 1, \dots, j_1; \quad k = 0, 1, 2, \dots, 2^j - 1, \\ \bar{\alpha}_{j_0k} &= \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \varphi_{j_0k}(X_i, Z_i)}{1 - G(Z_i)}, \quad k = 0, 1, 2, \dots, 2^{j_0} - 1, \\ \bar{\beta}_{jk} &= \frac{1}{n} \sum_{i=1}^n \frac{\delta_i \varphi_{jk}(X_i, Z_i)}{1 - G(Z_i)}, \quad j = j_0, j_0 + 1, \dots, j_1; \quad k = 0, 1, 2, \dots, 2^j - 1. \end{aligned}$$

Then the following equations hold.

$$\begin{aligned} \hat{\alpha}_{j_0k} &= \bar{\alpha}_{j_0k} + \bar{W}_{j_0k} + R_{n,j_0k}, & E(R_{n,j_0k}^2) &= O\left(\frac{1}{n^2}\right) \int \varphi_{j_0k}^2 dF^0, \\ \hat{\beta}_{jk} &= \bar{\beta}_{jk} + \bar{W}_{jk} + R_{n,jk}, & E(R_{n,jk}^2) &= O\left(\frac{1}{n^2}\right) \int \varphi_{jk}^2 dF^0, \end{aligned}$$

where

$$\begin{aligned} W_{j_0k}(Z_i) &= U_{j_0k}(Z_i) - V_{j_0k}(Z_i), & W_{jk}(Z_i) &= U_{jk}(Z_i) - V_{jk}(Z_i), \\ \bar{W}_{j_0k} &= \frac{1}{n} \sum_{i=1}^n W_{j_0k}(Z_i), & \bar{W}_{jk} &= \frac{1}{n} \sum_{i=1}^n W_{jk}(Z_i), \end{aligned}$$

and

$$\begin{aligned}
U_{j_0k}(Z_i) &= \frac{1 - \delta_i}{1 - H(Z_i)} \int_{Z_i}^{\tau_H} \varphi_{j_0k}(x, y) F^0(dx, dy), \\
U_{jk}(Z_i) &= \frac{1 - \delta_i}{1 - H(Z_i)} \int_{Z_i}^{\tau_H} \varphi_{jk}(x, y) F^0(dx, dy), \\
V_{j_0k}(Z_i) &= \int_{-\infty}^{\tau_H} \int_{-\infty}^{\tau_H} \frac{\varphi_{j_0k}(x, y) I(v < Z_i \wedge y)}{[1 - H(v)][1 - G(v)]} G(dv) F^0(dx, dy), \\
V_{jk}(Z_i) &= \int_{-\infty}^{\tau_H} \int_{-\infty}^{\tau_H} \frac{\varphi_{jk}(x, y) I(v < Z_i \wedge y)}{[1 - H(v)][1 - G(v)]} G(dv) F^0(dx, dy).
\end{aligned}$$

Proof of Lemma 3.1: The proof is analogous to that of Lemma 4.1 of Li (2003) for density estimation. In order to shorten it, we will skip the detailed lengthy proof and adopt facts from Stute (1993, 1995, 1996, 1999) and Li (2003), whenever it will be convenient. When there is no covariate, Stute (1995) considered convergence in distribution for the Kaplan-Meier integral $\int \varphi d\hat{F}_n$ and obtained the following result (Stute, 1995, p.434)

$$\begin{aligned}
\int \varphi(x) d\hat{F}_n(x) &= \int \varphi(w) \gamma_0(w) \tilde{H}_n^1(dw) + \iint \frac{I(v < w) \varphi(w) \gamma_0(w)}{1 - H(v)} \tilde{H}^1(dw) \tilde{H}_n^0(dv) \\
&\quad - \iiint \frac{I(v < u, v < w) \varphi(w) \gamma_0(w)}{[1 - H(v)]^2} \tilde{H}^0(dv) \tilde{H}^1(dw) H_n(du) + R_n,
\end{aligned} \tag{3.1}$$

where $|R_n| = o_p(n^{-1/2})$. For details on the definitions of γ_0 , \tilde{H}^0 , \tilde{H}_n^0 , \tilde{H}^1 , \tilde{H}_n^1 , etc, see Stute (1995). If we replace $\varphi(x)$ in (3.1) with $\varphi_{ij}(x) = \psi_{ij}(x) I(x \leq T)$, $T < \tau_H$, we obtain

$$\hat{b}_{ij} = \tilde{b}_{ij} + \bar{W}_{ij} + R_{n,ij}, \tag{3.2}$$

where

$$\begin{aligned}
\hat{b}_{ij} &= \int \varphi_{ij}(x) d\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^n \frac{\delta_k \varphi_{ij}(Z_k)}{1 - \hat{G}_n(Z_k-)}, \\
\tilde{b}_{ij} &= \int \varphi_{ij}(w) \gamma_0(w) \tilde{H}_n^1(dw) = \frac{1}{n} \sum_{k=1}^n \frac{\delta_k \varphi_{ij}(Z_k)}{1 - G(Z_k)}, \\
\bar{W}_{ij} &= \frac{1}{n} \sum_{k=1}^n W_{ij}(Z_k) = \frac{1}{n} \sum_{k=1}^n (U_{ij}(Z_k) - V_{ij}(Z_k)), \\
U_{ij}(Z_k) &= \frac{1 - \delta_k}{1 - H(Z_k)} \int_{Z_k}^{\tau_H} \varphi_{ij}(w) F(dw),
\end{aligned}$$

etc. For details on the other terms and proof, see Lemma 4.1 in Li (2003). Identity (3.2) approximates the empirical wavelet coefficients (\hat{b}_{ij}) with an average of i.i.d. random variables $(\tilde{b}_{ij} + \bar{W}_{ij})$ and a sufficiently small error $R_{n,ij}$. Stute (1995) shows $|R_n| =$

$o_p(n^{-1/2})$ for his central limit theorem. Since we consider here the mean integrated square error measure of the estimator, we need to control the second moment of the reminder term $R_{n,ij}$. Lemma 4.1 in Li (2003) shows that \tilde{b}_{ij} and \overline{W}_{ij} are dominating terms and $R_{n,ij}$ is a negligible term such that $ER_{n,ij}^2 = O(n^{-2}) \int \varphi_{ij}^2 dF$. Based on this approximation, the empirical wavelet coefficients \hat{b}_{ij} basically can be treated as \tilde{b}_{ij} , which is an average of i.i.d. random variables.

When a covariate is present, Stute (1993) extends $\hat{F}_n(x)$ to obtain an estimator $\hat{F}_n^0(x, y)$ in (2.5) of the joint distribution function $F^0(x, y)$ of (X, Y) when Y is subject to censoring and X is observable and derives an analogous result as follows:

$$\begin{aligned} \int \varphi(x, y) \hat{F}_n^0(dx, dy) &= \int \varphi(x, w) \gamma_0(w) \tilde{H}_n^{11}(dx, dw) \\ &+ \iint \frac{I(v < w) \varphi(x, w) \gamma_0(w)}{1 - H(v)} \tilde{H}^{11}(dx, dw) \tilde{H}_n^0(dv) \\ &- \iiint \frac{I(v < u, v < w) \varphi(x, w) \gamma_0(w)}{[1 - H(v)]^2} \tilde{H}^0(dv) \tilde{H}^{11}(dx, dw) H_n(du) + R_n, \end{aligned} \quad (3.3)$$

where $|R_n| = o_p(n^{-1/2})$. For details on the definitions of \tilde{H}^0 , \tilde{H}_n^0 , \tilde{H}^{11} , \tilde{H}_n^{11} and other terms, see Stute (1996, p.463) and Stute (1999, p.1095). If we replace $\varphi(x, y)$ in (3.3) with $\varphi_{jk}(x, y) = y\psi_{jk}(x)$, we obtain

$$\hat{\beta}_{jk} = \bar{\beta}_{jk} + \overline{W}_{jk} + R_{n,jk},$$

where

$$\begin{aligned} \hat{\beta}_{jk} &= \int \varphi_{jk}(x, y) \hat{F}_n^0(dx, dy) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i Z_i \psi_{jk}(X_i)}{1 - \hat{G}_n(Z_i-)}, \\ \bar{\beta}_{jk} &= \int \varphi_{ij}(x, w) \gamma_0(w) \tilde{H}_n^{11}(dx, dw) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i Z_i \psi_{jk}(X_i)}{1 - G(Z_i)}, \end{aligned}$$

etc. Based on our assumption, $Y \leq B$ and $1 - G(B) > 0$, we have $Z = \min\{Y, T\} \leq B$. Thus all the denominators in $\bar{\beta}_{jk}$, \overline{W}_{jk} , and $R_{n,jk}$ are bounded away from zero. Following the proof of Lemma 4.1 in Li (2003) and noting that $\hat{\alpha}_{j_0k}$, $\bar{\alpha}_{j_0k}$, $\hat{\beta}_{jk}$, $\bar{\beta}_{jk}$ and 2^{j_0} here play roles of \hat{b}_j , \tilde{b}_j , \hat{b}_{ij} , \tilde{b}_{ij} and p there (similarly for U, V and W), we can prove $E(R_{n,jk}^2) = O(n^{-2}) \int \varphi_{jk}^2 dF^0$. \square

Once we establish the above approximation for the empirical wavelet coefficients $\hat{\beta}_{jk}$ as an average of i.i.d. random variables, we are ready to provide the large deviation result on $\hat{\beta}_{jk}$. The following lemma is needed for large deviation behavior on the empirical wavelet coefficients $\hat{\beta}_{jk}$.

Lemma 3.2. (*Bennett's or Bernstein's Inequality*) (Härdle, et al., 1998, p243) Let X_1, X_2, \dots, X_n be independent random variables such that $E(X_i) = 0$, $|X_i| \leq M$, $b_n^2 = \sum_{i=1}^n E(X_i^2)$. Then for any $\lambda \geq 0$,

$$P\left(\left|\sum_{i=1}^n X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2(b_n^2 + \frac{\lambda M}{3})}\right), \quad \forall \lambda \geq 0.$$

The following large deviation result is analogous to that of Lemma 4.4 in Li (2004) for density estimation. The difference is that here because of the covariate we deal with a triple of variables, instead of the bivariate case. Since the estimators are different, the threshold constants are different also. The lemma basically says that there is a negligible probability that the empirical coefficients and theoretical coefficients differ greatly. In the proofs below, C represents a generic finite constant, the value of which may change from line to line in the sequel.

Lemma 3.3. Let $\hat{\beta}_{jk}$ be defined as in equation (2.4). Then

$$P\left(\left|\hat{\beta}_{jk} - \beta_{jk}\right| > d\sqrt{\frac{\ln n}{n}}\right) = O(n^{-1}),$$

for all $j \in [j_0, j_1]$, $2^{j_0} \simeq \log n$, $2^{j_1} \simeq n(\log_2 n)^{-2}$, and $k = 0, 1, \dots, 2^j - 1$.

Proof of Lemma 3.3: For any positive numbers r_1, r_2 and r_3 , such that $r_1 + r_2 + r_3 = 1$, from Lemma 3.1, we have

$$\begin{aligned} P\left(\left|\hat{\beta}_{jk} - \beta_{jk}\right| > d\sqrt{\frac{\ln n}{n}}\right) &\leq P\left(\left|\bar{\beta}_{jk} - \beta_{jk}\right| > r_1 d\sqrt{\frac{\ln n}{n}}\right) \\ &\quad + P\left(\left|\bar{W}_{jk}\right| > r_2 d\sqrt{\frac{\ln n}{n}}\right) \\ &\quad + P\left(\left|R_{n,jk}\right| > r_3 d\sqrt{\frac{\ln n}{n}}\right) \\ &=: P_1 + P_2 + P_3. \end{aligned}$$

We want to show that $P_i = O(n^{-1})$, $i = 1, 2$ and 3 . From Lemma 3.1, we can write $\bar{\beta}_{jk} - \beta_{jk} = n^{-1} \sum_i \xi_{jk,i}$, where $\xi_{jk,i} = \delta_i Z_i \psi_{jk}(X_i)(1 - G(Z_i))^{-1} - \beta_{jk}$. Note that $|\beta_{jk}| \leq \int |g| |\psi_{jk}(x)| dx \leq A2^{j/2} \sup |\psi|$. Thus we have $|\xi_{jk,i}| \leq B2^{j/2} \sup |\psi|(1 - G(B))^{-1} + A2^{j/2} \sup |\psi| = C2^{j/2}$. By direct calculation, $E(\xi_{jk,i}) = 0$, and

$$\text{Var}(\xi_{jk,i}) = E(\xi_{jk,i}^2) \leq E\{\delta_i^2 Z_i^2 \psi_{jk}(X_i)^2 (1 - G(Z_i))^{-2}\} \leq B^2(1 - G(B))^{-1}.$$

Hence, we can apply Lemma 3.2 with $\lambda = nr_1 d \sqrt{n^{-1} \ln n}$, $M = C2^{j/2}$, $b_n^2 = nB^2(1 - G(B))^{-1}$. Thus we have

$$P_1 \leq 2 \exp \left\{ - \frac{r_1^2 d^2 \ln n}{2(B^2(1 - G(B))^{-1} + 3^{-1} C r_1 d \sqrt{n^{-1} \ln n} 2^{j/2})} \right\}.$$

Based on our choice $2^{j_1} \simeq n(\log_2 n)^{-2}$, we have $\sqrt{n^{-1} \ln n} 2^{j/2} \rightarrow 0$, $\forall j \in [j_0, j_1]$. Hence we obtain, for all $\epsilon > 0$,

$$P_1 \leq 2 \exp \left\{ - \frac{1}{2} (1 - \epsilon) B^{-2} (1 - G(B)) r_1^2 d^2 \ln n \right\}.$$

Choose ϵ sufficiently small and r_1 sufficiently large and close to 1. Based on our choice of d , we have $P_1 = O(n^{-1})$, $\forall j \in [j_0, j_1]$ and k . Applying the same argument as in Lemma 4.4 in Li (2004), we can show $P_2 = P_3 = O(n^{-1})$ also. \square

Now we are ready to provide the outline of the proof of Theorem 3.1.

Proof of Theorem 3.1: We can break the proof of Theorem 3.1 into several parts. The orthogonality of ϕ and ψ implies that

$$E \|\hat{g} - g\|_2^2 = I_1 + I_2 + I_3 + I_4,$$

where

$$\begin{aligned} I_1 &= \sum_{k=0}^{2^{j_0}-1} E(\hat{\alpha}_{j_0 k} - \alpha_{j_0 k})^2, & I_2 &= \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \beta_{jk})^2, \\ I_3 &= \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \beta_{jk})^2, & I_4 &= \sum_{j=j_1+1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{jk}^2. \end{aligned}$$

Here $\hat{\theta}_{jk} = \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > d \sqrt{n^{-1} \ln n})$ and $d = \sqrt{2(1 - G(B))^{-1}} B$. The smoothing parameter j_α is carefully chosen such that $2^{j_\alpha} \simeq (n(\log_2 n)^{-1})^{1/(1+2\alpha)}$ in order to balance the two terms I_2 and I_3 . In order to complete the proof, it suffices to show that $I_1 = o((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)})$, $I_2 \leq C(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}$, $I_3 \leq C(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}$ and $I_4 = o((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)})$. Their proofs are very similar to Lemmas 4.5-4.8 in Li (2004) for the density estimation case. In order to make the proof complete, we provide their proofs in the appendix. \square

Remark 3.1. The above term-by-term hard thresholded wavelet estimator defined as in (2.3) and (2.4) is adaptive in the sense that it does not depend on unknown parameters

α , p and q . Minimax theory indicates that the best convergence rate over $\mathcal{G}_{pq}^\alpha(M)$ is $n^{-2\alpha/(2\alpha+1)}$. Thus, the above estimator achieves optimal convergence rates up to a logarithmic term, without knowing *a priori* the smoothness parameters. In the case $p < 2$, Donoho *et al.* (1996) shows that the traditional linear estimator cannot achieve the rates stated in Theorem 3.1. Hence non-linear wavelet estimators typically achieve better convergence rates over large function spaces. The same result holds for an analogous soft thresholding wavelet estimator. We conjecture that a block thresholded estimator similar to that in Hall *et al.* (1998) or Cai (1999) can be constructed so that it attains exact minimax convergence rates without the logarithmic penalty. The proof would likely follow the arguments of Hall *et al.* (1998), but it would be too lengthy to discuss these details here.

Remark 3.2. The estimator defined by (2.3) and (2.4) is analogous to that of Hall and Patil (1996a) for the complete data case. Without censoring, the two estimators would be the same. The choices of the primary resolution level j_0 and upper level j_1 here are not unique. For more on the choice of smoothing parameters and threshold, see Hall and Patil (1996a, 1996b).

Remark 3.3. Our estimator \hat{g} in (2.3) corresponds to Koul's transformation (Koul *et al.*, 1981), *i.e.*, replacing observation Z_i with $Y_i^* = \delta_i Z_i [1 - \hat{G}_n(Z_i -)]^{-1}$. There are other transformations available for the censored data, for example, Buckley and James (1979) (henceforth *BJ*), Fan and Gijbels (1994), McKeague *et al.* (2001), among others. The *BJ* transformation replaces Z_i with its conditional expectation $Y_i^* = E(Y|Y > Z_i, X_i, \delta_i)$. Fan and Gijbel (1994) show that the *BJ* transformation has smaller variability than Koul's transformation. In the context of the conditional independence of Y and T given the covariate X , we can propose the following corresponding wavelet estimator based on the *BJ* transformation:

$$\tilde{g}(x) = \sum_{k=0}^{2^{j_0}-1} \tilde{\alpha}_{j_0 k} \phi_{j_0 k}(x) + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \tilde{\beta}_{j k} I(|\tilde{\beta}_{j k}| > \tilde{d} \sqrt{n^{-1} \ln n}) \psi_{j k}(x),$$

where the empirical wavelet coefficients are defined

$$\begin{aligned} \tilde{\alpha}_{j_0 k} &= \frac{1}{n} \sum_{i=1}^n \left\{ Z_i \delta_i + \left[Z_i + \frac{\int_{Z_i}^{\infty} \hat{S}_n(y, X_i) dy}{\hat{S}_n(Z_i, X_i)} \right] (1 - \delta_i) \right\} \phi_{j_0 k}(X_i), \\ \tilde{\beta}_{j k} &= \frac{1}{n} \sum_{i=1}^n \left\{ Z_i \delta_i + \left[Z_i + \frac{\int_{Z_i}^{\infty} \hat{S}_n(y, X_i) dy}{\hat{S}_n(Z_i, X_i)} \right] (1 - \delta_i) \right\} \psi_{j k}(X_i). \end{aligned}$$

Here $\hat{S}_n(y, x)$ is the local Kaplan-Meier estimator of the conditional survival function of Y given X . Upon choosing proper smoothing parameters j_0, j_1 and threshold \tilde{d} , we would expect the above estimator \tilde{g} to behave better than \hat{g} and conjecture that we can obtain convergence rates similar to Theorem 3.1. The proof would likely follow that of Theorem 3.1, but it would be too involved to discuss these details here.

4 Simulation

To investigate the performance of the proposed wavelet estimator, we present a modest simulation study. Since Fan and Gijbels (1994) have shown that the BJ transformation has smaller variability than Koul's transformation, we have used a wavelet estimator based on the BJ transformation for comparison purposes in this simulation. However, since the BJ transformation depends on the unknown regression, the computation of the empirical wavelet coefficients would involve an iterative algorithm which is much too unwieldy to use in practice. Therefore, we propose the use of the explicit local average transformation provided by Fan and Gijbel (1994, p.562). The transformation replaces the censored observation Z_i with Y_i^* , which is a weighted average of all uncensored responses which are larger than Z_i within a small neighborhood of X_i , i.e.,

$$Y_i^* = \frac{\sum_{j:Z_j>Z_i} Z_j K\left(\frac{X_i-X_j}{(X_{i+k}-X_{i-k})/2}\right) \delta_j}{\sum_{j:Z_j>Z_i} K\left(\frac{X_i-X_j}{(X_{i+k}-X_{i-k})/2}\right) \delta_j},$$

where K is a nonnegative kernel function and k plays the role of bandwidth. The value of k can be determined by *cross-validation*.

In order to compare our wavelet estimator to Fan and Gijbel's local linear regression smoother (henceforth local linear), we use the same functions in the simulation study as in Fan and Gijbel (1994), i.e., $Y_i = 4.5 - 64X_i^2(1 - X_i)^2 - 16(X_i - 0.5)^2 + 0.25\epsilon_i$, $\epsilon_i \sim i.i.d. N(0, 1)$. For the convenience of the discrete wavelet transform, we let $X_i = i/n, i = 1, 2, \dots, n$, where n is the sample size. We consider three different sample sizes: $n = 256, 512$ and 1024 . As in Fan and Gijbels (1994), the censoring time T_i is conditionally independent of the survival time Y_i given X_i and is distributed as $(T_i|X_i = x) \sim exp(t(x))$, where $t(x)$ is the mean conditional censoring time given by

$$t(x) = \begin{cases} 3(1.25 - |4x - 1|), & \text{if } 0 \leq x \leq 0.5; \\ 3(1.25 - |4x - 3|), & \text{if } 0.5 < x \leq 1. \end{cases}$$

For the above censoring variable, approximately 40% of the data are censored. We also consider another censoring variable $(T_i|X_i = x) \sim \exp(2.2 * t(x))$ which results in approximately 20% of data are censored. For numerical comparisons we consider the average norm (ANorm) of the estimators at the sample points

$$ANorm = \frac{1}{N} \sum_{l=1}^N \left(\sum_{i=1}^n (\hat{f}_l(x_i) - f(x_i))^2 \right)^{1/2},$$

where \hat{f}_l is the estimate of f in l -th replication and N is the total number of replications. Since different wavelets yield very similar results, we only use Daubechies's compactly support wavelet *Symmlet 8*. Using the idea of *cross-validation*, we select the threshold based on the minimum value of the average norm. The associated estimator with this minimum turns out to be very close to the Stein Unbiased Risk Estimator (Donoho *et al.* (1995)). The simulation results for different sample sizes and different censoring proportion are summarized in Table 1. Based on these results, our wavelet estimator has

Table 1: Average Norm from $N = 100$ replications.

	20% censored			40% censored		
	n=256	n=512	n=1024	n=256	n=512	n=1024
Local Linear	3.844	5.549	7.886	3.869	5.497	7.860
Wavelet	3.867	5.519	7.774	3.850	5.463	7.780

a very similar average norm which is usually slightly smaller than that of the local linear smoother.

The second example we considered is the following model: $Y_i = g(X_i) + \epsilon_i$, where $\epsilon_i \sim i.i.d. N(0, 1)$ $i = 1, \dots, n$, ($n = 256, 512$), and $g(x)$ is a piece-wise HeaviSine function:

$$g(x) = \begin{cases} 4\sin(4\pi x) + 20, & \text{if } 0 \leq x < 0.3; \\ 4\sin(4\pi x) + 18, & \text{if } 0.3 \leq x < 0.7; \\ 4\sin(4\pi x) + 20, & \text{if } 0.7 \leq x \leq 1; \end{cases}$$

We also considered the censoring time T_i to be conditionally independent of the survival time Y_i given X_i and to be distributed as $(T_i|X_i = x) \sim \exp(t(x))$, where $t(x) = 4g(x)$ results in approximately 40% censoring and $t(x) = 2g(x)$ results in approximately 20%

censoring. The average norm for two estimators for different sample sizes and censoring are summarized in Table 2. Based on the above simulation results, we found that for

Table 2: Average Norm from $N = 100$ replications.

	20% censored		40% censored	
	n=256	n=512	n=256	n=512
Local Linear	7.691	9.199	8.540	10.721
Wavelet	7.332	8.941	9.421	11.851

light censoring (approximately 20%) and moderate sample sizes, our wavelet estimator does slightly better than Fan and Gijbel’s local linear estimator. However, for heavier censoring, this advantage is lost.

5 Data Analysis

In this section, we apply our method to the Stanford Heart Transplant data, which has previously been analyzed by Fan and Gijbel (1994) and others. (For more information about this data set, see Miller and Halpern (1982)). This data consists of 184 patients who received a heart transplant between October 1967 and February 1980. Of these, 119 died during the follow-up period and 65 (those patients who lived beyond February 1980) were censored. We consider all 184 cases, their survival times which are log-transformed, as well as their age at transplant. The following figure includes two estimators: the smooth one is the local linear estimator, while the rough one is wavelet estimator. We can see these two estimators are very close. The wavelet estimator is obviously not as smooth as the local linear one found in Fan and Gijbels (1994). However, it clearly shows the same trend. It is evident that, because of the adaptability of the wavelet estimator to many different type of non-smoothness, a price will be paid on the estimation of a truly smooth curve. However, in this case whether or not the true curve is smooth is unknown. It is possible that the Fan-Gijbels method oversmooths and/or the wavelet method undersmooths.

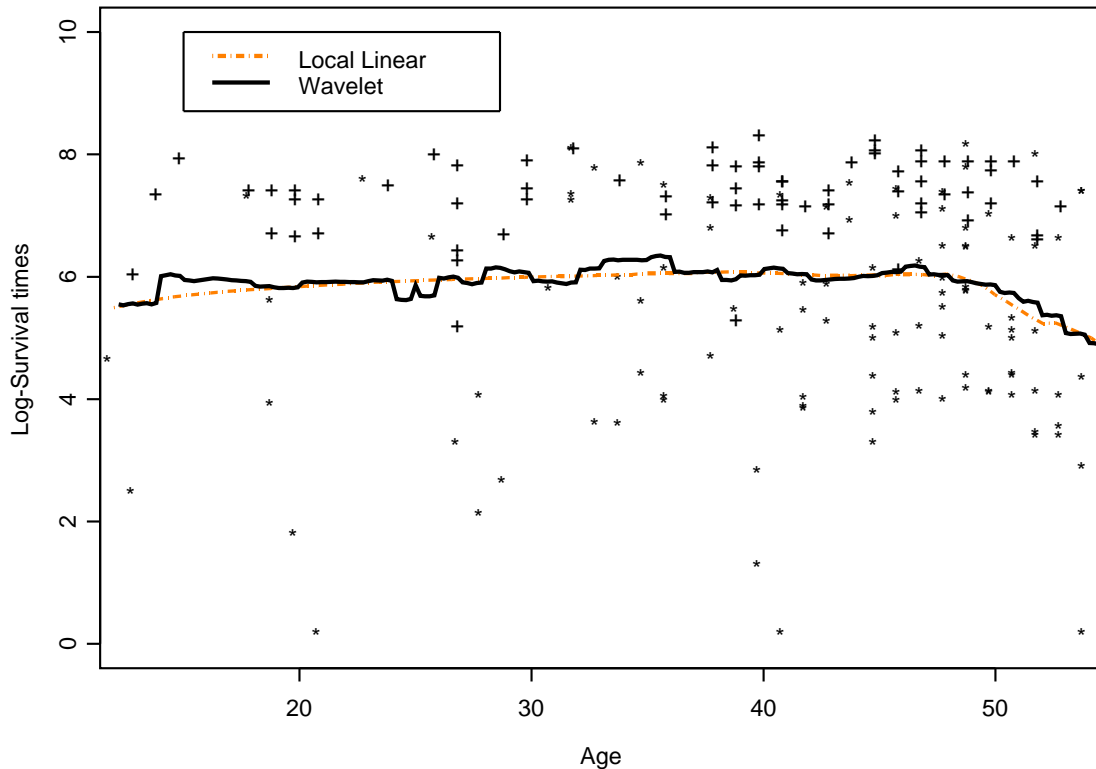


Figure 1: Stanford Heart Transplant data with log-survival time plotted against age. The symbol + indicates survival time is censored and symbol * indicates survival time is observed. Two estimators are plotted.

6 Conclusion

For the problem of nonparametric regression function estimation with randomly right censored data, we have shown that wavelet estimators based on thresholding the empirical wavelet coefficients can achieve nearly optimal rates within logarithmic terms over a large range of Besov function classes B_{pq}^α , $\alpha > 1/p$, $p \geq 1$, $q \geq 1$. In general, wavelets outperform other methods when analysing functions with non-smooth features or discontinuities. For randomly right censored regression problems wavelet analysis, when combined with the proper data transformation such as those used by Fan and Gijbels (1994), provides results that are consistent with those using local linear smoothers. It is perhaps surprising

that the wavelet estimators did not perform significantly better with heavier censoring than these smoothers in the simulation studies but the results could possibly be improved by using different types of wavelet coefficient thresholding such as a block thresholded estimator analogous to that of Hall *et al.* (1998) or Cai (1999). Another approach would be to use coefficient dependent thresholding as advocated by Kovac and Silverman (2000) and Von Sachs and MacGibbon (2000). These are interesting subjects to pursue in further studies.

Acknowledgements The authors wish to thank Professors Jianqing Fan and Irène Gijbels for making their program code used for the local linear regression smoother available to us for this research. The second author wishes to thank NSERC of Canada for the partial support of this research.

REFERENCES

- Buckley, J. and James, I. R. (1979). Linear regression with censored data. *Biometrika* **66**, 429-436.
- Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898-924.
- Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harm. Anal.* **1**, 54-82.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.
- Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425-455.
- Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinking. *J. Am. Statist. Assoc.* **90**, 1200-1224.
- Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 301-369.
- Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508-539.

- Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Am. Statist. Assoc.* **89**, 560–570.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.
- Hall, P., Kerkycharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26**, 922–942.
- Hall, P., Kerkycharian, G. and Picard, D. (1999a). On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica* **9**, 33–50.
- Hall, P., Kerkycharian, G. and Picard, D. (1999b). A note on the wavelet oracle. *Statistics and Probability Letters* **43**, 415–420.
- Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of non-linear wavelet-based density estimators. *Ann. Statist.* **23**, 905–928.
- Hall, P. and Patil, P. (1996a). On the choice of smoothing parameter, threshold and truncation in nonparametric regression by nonlinear wavelet methods. *J. Roy. Statist. Soc. Ser. B* **58**, 361–377.
- Hall, P. and Patil, P. (1996b). Effect of threshold rules on performance of wavelet-based curve estimators. *Statistic Sinica* **6**, 331–345.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Boston: Cambridge University Press.
- Härdle, W., Kerkycharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications. Lecture Notes in Statistics* **129** Springer, NewYork.
- Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9**, 1276–1288.
- Kovac, A. and Silverman, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Am. Statist. Assoc.* **95**, 172–183.
- Li, L. (2003). Non-linear wavelet-based density estimators under random censorship. *J. Statistical Planning and Inference* **117**, 35–58.
- Li, L. (2004). On the minimax optimality of wavelet estimators with censored data. *Submitted*.

- McKeague, I. W., Subramanian, S. and Sun, Y. (2001). Median regression and the missing information principle. *J. Nonparam. Statist.* **13**, 709-727.
- Meyer, Y. (1992). *Wavelets and Operators*, Cambridge University Press, Cambridge.
- Miller, R. G. (1976). Least squares regression with censored data. *Biometrika* **63**, 449-464.
- Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521-531.
- Portnoy, S. (2003). Censored Regression Quantiles. *J. Am. Statist. Assoc.* **98**, 1001-1012.
- Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multiv. Analysis* **45**, 89-103.
- Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.* **23**, 422-439.
- Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.* **23**, 461-471.
- Stute, W. (1999). Nonlinear censored regression. *Statistica Sinica* **9**, 1089-1102.
- Triebel, H. (1992). *Theory of Function Spaces II*. Birkhäuser, Basel.
- Von Sachs, R. and MacGibbon, B. (2000). Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statistics* **27**, 475-499.
- Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley & Sons, Inc.

A Appendix

First we need the following norm inequality to bound the term I_4 in Lemma A.3. Its proof is straight forward and therefore omitted.

Lemma A.1. *Let $u \in \mathbb{R}^n$, and $0 < p_1 \leq p_2 \leq \infty$. Then the following inequalities hold:*

$$\|u\|_{p_2} \leq \|u\|_{p_1} \leq n^{\frac{1}{p_1} - \frac{1}{p_2}} \|u\|_{p_2}.$$

Lemma A.2. Let $\hat{\alpha}_{j_0k}$ be defined as in equation (2.4). Then

$$I_1 = \sum_{k=0}^{2^{j_0}-1} E(\hat{\alpha}_{j_0k} - \alpha_{j_0k})^2 = o\left((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}\right).$$

Proof. From the approximation to $\hat{\alpha}_{j_0k}$ in Lemma 3.1 and an elementary inequality, we have

$$\begin{aligned} I_1 &\leq 3 \left[\sum_{k=0}^{2^{j_0}-1} E(\bar{\alpha}_{j_0k} - \alpha_{j_0k})^2 + \sum_{k=0}^{2^{j_0}-1} E\bar{W}_{j_0k}^2 + \sum_{k=0}^{2^{j_0}-1} ER_{n,j_0k}^2 \right] \\ &=: 3(I_{11} + I_{12} + I_{13}). \end{aligned}$$

Applying the same arguments as in Lemma 4.2 in Li (2003, p42-43) and noticing that the primary level 2^{j_0} plays the role of p in Li (2003), we can have

$$I_{11} = O(n^{-1}2^{j_0}), \quad I_{12} = o(n^{-1}2^{j_0}), \quad \text{and} \quad I_{13} = O(n^{-2}2^{j_0}).$$

Based on our choice j_0 with $2^{j_0} \simeq \log_2 n$, the lemma is proved. \square

Lemma A.3. Let β_{jk} be defined as in expansion (2.2). Then

$$I_4 = \sum_{j=j_1+1}^{\infty} \sum_{k=0}^{2^j-1} \beta_{jk}^2 = o\left((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}\right).$$

Proof. From the wavelet expansion (2.2), the wavelet coefficients $\beta_{jk} = \int g(x)\psi_{jk}(x)dx$. Because $\text{supp}g \subset [0, 1]$ and $\text{supp}\psi \subset [-v, v]$, we have, for any level j , there are at most $C2^j$ non-zero coefficients β_{jk} 's. In order to bound the above term I_4 , we use the equivalence of the Besov norm and the wavelet coefficients sequence norm in (2.1). For this purpose, we need to separate the cases $p \leq 2$ and $p > 2$. First, let's consider $p \leq 2$. From Lemma A.1 and (2.1), we have $\|\beta_j\|_2 \leq \|\beta_j\|_p \leq M2^{-js}$. Thus $\sum_{k=0}^{2^j-1} \beta_{jk}^2 \leq M^22^{-2js}$. Since $\alpha p > 1$ and $s > 1/2$, we have $I_4 \leq \sum_{j=j_1+1}^{\infty} M^22^{-2js} = M^22^{-2j_1s}2^{-2s}(1 - 2^{-2s})^{-1}$. Based on our choice j_1 with $2^{j_1} \simeq n(\log_2 n)^{-2}$ and $2s = 1 + 2(\alpha - 1/p) > 2\alpha/(2\alpha + 1)$, we obtain $I_4 = o\left((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}\right)$. For $p > 2$, from Lemma A.1, we have $\|\beta_j\|_2 \leq (C2^j)^{\frac{1}{2}-\frac{1}{p}}\|\beta_j\|_p \leq C2^{-j\alpha}$. Thus we have

$$I_4 \leq C \sum_{j_1+1}^{\infty} 2^{-2j\alpha} = C2^{-2j_1\alpha}2^{-2\alpha}(1 - 2^{-2\alpha})^{-1} = C(n^{-1}(\log_2 n)^2)^{2\alpha}.$$

Since $\alpha > 0$, we have $I_4 = o\left((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}\right)$. Together with $p \leq 2$, this completes the proof of the lemma. \square

Lemma A.4. Let $\hat{\beta}_{jk}$ be defined as in equation (2.4) and $\hat{\theta}_{jk} = \hat{\beta}_{jk}I(|\hat{\beta}_{jk}| > d\sqrt{n^{-1}\ln n})$.

Then

$$I_2 = \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \beta_{jk})^2 \leq C(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)},$$

where $j_\alpha = j_\alpha(n)$, such that $2^{j_\alpha} \simeq (n^{-1} \log_2 n)^{1/(1+2\alpha)}$.

Proof. We write

$$\begin{aligned} I_2 &\leq 2 \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} E \left[\beta_{jk}^2 I(|\hat{\beta}_{jk}| \leq d\sqrt{n^{-1}\ln n}) \right] \\ &\quad + 2 \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} E \left[(\hat{\beta}_{jk} - \beta_{jk})^2 I(|\hat{\beta}_{jk}| > d\sqrt{n^{-1}\ln n}) \right] \\ &=: 2(I_{21} + I_{22}). \end{aligned}$$

As for the first term I_{21} , we write

$$\begin{aligned} I_{21} &\leq \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} \beta_{jk}^2 I(|\beta_{jk}| \leq 2d\sqrt{n^{-1}\ln n}) \\ &\quad + \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} \beta_{jk}^2 P(|\hat{\beta}_{jk} - \beta_{jk}| > d\sqrt{n^{-1}\ln n}) \\ &=: I_{211} + I_{212}. \end{aligned}$$

Since there are at most $C2^j$ non-zero terms of β_{jk} 's for each level j , we have

$$I_{211} \leq \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} 4d^2 \frac{\ln n}{n} \leq C \frac{\log_2 n}{n} \sum_{j=j_0}^{j_\alpha} 2^j \leq C(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}.$$

As for term I_{212} , applying the large deviation result in Lemma 3.3, we have $I_{212} \leq Cn^{-1} \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} \beta_{jk}^2$. Applying the same argument as in Lemma A.3 for both $p \leq 2$ and $p > 2$, it is easy to obtain $I_{212} = o((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)})$. Therefore we obtain $I_{21} \leq C(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}$. As for term I_{22} , from the approximation of $\hat{\beta}_{jk}$ in Lemma 3.1, we have

$$\begin{aligned} I_{22} &\leq 3 \left[\sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} E(\bar{\beta}_{jk} - \beta_{jk})^2 + \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} E\bar{W}_{jk}^2 + \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} ER_{n,jk}^2 \right] \\ &=: 3(I_{221} + I_{222} + I_{223}). \end{aligned}$$

By direct calculation, we can obtain $E(\bar{\beta}_{jk} - \beta_{jk})^2 = O(n^{-1})$. Since there are at most $C2^j$ non-zero terms for each j , we have $I_{221} \leq \sum_{j=j_0}^{j_\alpha} Cn^{-1}2^j = Cn^{-1}(2^{j_\alpha-j_0}) \leq Cn^{-1}(n(\log_2 n)^{-1})^{1/(1+2\alpha)} = C(\log_2 n)^{-1}(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)} = o((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)})$. By a computation analogous to that in Lemma 4.2 in Li (2003), we can obtain $EU_{jk}^2(Z_1) = EV_{jk}^2(Z_1) = O(1)$. Noticing that there are at most $C2^j$ non-zero terms of U_{jk} 's and V_{jk} 's for each j , we have $I_{222} = \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} n^{-1}EW_{jk}^2(Z_1) \leq \sum_{j=j_0}^{j_\alpha} Cn^{-1}2^j$. By the exact same argument as for I_{221} , we have $I_{222} = o((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)})$. For term I_{223} , from lemma 3.1, we have

$$I_{223} = \sum_{j=j_0}^{j_\alpha} \sum_{k=0}^{2^j-1} \frac{C}{n^2} \int \varphi_{jk}^2 dF^0 \leq \frac{C}{n^2} \sum_{j=j_0}^{j_\alpha} 2^{2j} \leq \frac{C}{n^2} 2^{2j_\alpha} = o\left((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}\right).$$

Together with I_{221} and I_{222} , this completes the proof of the Lemma. \square

Lemma A.5. Let $\hat{\beta}_{jk}$ be defined as in equation (2.4) and $\hat{\theta}_{jk} = \hat{\beta}_{jk}I(|\hat{\beta}_{jk}| > d\sqrt{n^{-1} \ln n})$.

Then

$$I_3 = \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} E(\hat{\theta}_{jk} - \beta_{jk})^2 \leq C(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}.$$

Proof. As in Lemma A.4, we have

$$\begin{aligned} I_3 &\leq 2 \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} E\left[\beta_{jk}^2 I(|\hat{\beta}_{jk}| \leq d\sqrt{n^{-1} \ln n})\right] \\ &\quad + 2 \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} E\left[(\hat{\beta}_{jk} - \beta_{jk})^2 I(|\hat{\beta}_{jk}| > d\sqrt{n^{-1} \ln n})\right] \\ &=: 2(I_{31} + I_{32}). \end{aligned}$$

For the first term I_{31} , we have

$$\begin{aligned} I_{31} &\leq \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 I(|\beta_{jk}| \leq 2d\sqrt{n^{-1} \ln n}) \\ &\quad + \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 P(|\hat{\beta}_{jk} - \beta_{jk}| > d\sqrt{n^{-1} \ln n}) \\ &=: I_{311} + I_{312}. \end{aligned}$$

In order to deal with term I_{311} , we need to separate $p \geq 2$ and $p < 2$. Let's first consider $p \geq 2$. Applying the same argument as in Lemma A.3, we have

$$I_{311} \leq \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 \leq C \sum_{j=j_\alpha+1}^{j_1} 2^{-2j\alpha} \leq C2^{-2\alpha j_\alpha} \leq C(n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}.$$

For the case $p > 2$, we have

$$\begin{aligned} I_{311} &= \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 I(\beta_{jk}^2 \leq 4d^2 \frac{\ln n}{n}) \leq \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 \wedge 4d^2 \frac{\ln n}{n} \\ &= \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2 \wedge 4d^2 C \frac{\log_2 n}{n}, \end{aligned}$$

where $a \wedge b = \min\{a, b\}$. Noticing that $(1+2\alpha)^{-1} < p/2$, for all α and p , we can choose a positive number $t \in (0, 1)$, such that $(1+2\alpha)^{-1} < t < p/2$. Therefore we have

$$\begin{aligned} I_{311} &\leq \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \left(\beta_{jk}^2 \wedge C \frac{\log_2 n}{n} \right)^t \cdot \left(\beta_{jk}^2 \wedge C \frac{\log_2 n}{n} \right)^{1-t} \\ &\leq \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^{2t} \cdot \left(C \frac{\log_2 n}{n} \right)^{1-t} \\ &\leq C \left(\frac{\log_2 n}{n} \right)^{1-t} \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^{2t}. \end{aligned}$$

Since $2t < p$, from Lemma A.1, we have

$$\|\beta_j\|_{2t} \leq C(2^j)^{\frac{1}{2t}-\frac{1}{p}} \|\beta_j\|_p \leq C(2^j)^{\frac{1}{2t}-\frac{1}{p}} M 2^{-js}.$$

Hence,

$$\begin{aligned} I_{311} &\leq C \left(\frac{\log_2 n}{n} \right)^{1-t} \sum_{j=j_\alpha+1}^{j_1} M^{2t} 2^{-2jts} 2^{j(\frac{1}{2t}-\frac{1}{p})2t} \\ &= C M^{2t} \left(\frac{\log_2 n}{n} \right)^{1-t} \sum_{j=j_\alpha+1}^{j_1} 2^{j(1-t-2t\alpha)} \\ &= C \left(\frac{\log_2 n}{n} \right)^{2t\alpha} = o\left((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)} \right), \end{aligned}$$

where the last equality follows from $(1+2\alpha)^{-1} < t$. For the term I_{312} , from Lemma 3.3, we have $I_{312} \leq C n^{-1} \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^2$. Analogous to term I_{212} , using the same argument as in Lemma A.3, we can obtain $I_{312} = o((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)})$. For the term I_{32} , for any

$\eta \in (0, 1)$, we have

$$\begin{aligned}
I_{32} &\leq \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} E \left[(\hat{\beta}_{jk} - \beta_{jk})^2 I(|\beta_{jk}| > \eta d \sqrt{n^{-1} \ln n}) \right] \\
&\quad + \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} E \left[(\hat{\beta}_{jk} - \beta_{jk})^2 I(|\hat{\beta}_{jk} - \beta_{jk}| > (1-\eta) d \sqrt{n^{-1} \ln n}) \right] \\
&=: I_{321} + I_{322}.
\end{aligned}$$

Let's consider I_{321} first. From Lemma 3.1, applying the same argument as in I_{22} , we can obtain $E(\hat{\beta}_{jk} - \beta_{jk})^2 = O(n^{-1})$. Since $\beta_{jk}^2 > \eta^2 d^2 C \frac{\log_2 n}{n}$ in I_{321} , we have

$$\begin{aligned}
I_{321} &\leq \frac{C}{n} \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \left(\beta_{jk}^2 \frac{n}{\eta^2 d^2 C \log_2 n} \right)^{p/2} = \frac{C n^{p/2-1}}{(\log_2 n)^{p/2}} \sum_{j=j_\alpha+1}^{j_1} \sum_{k=0}^{2^j-1} \beta_{jk}^p \\
&\leq \frac{C n^{p/2-1}}{(\log_2 n)^{p/2}} \sum_{j=j_\alpha+1}^{j_1} M^p 2^{-j s p} = \frac{C M^p n^{p/2-1}}{(\log_2 n)^{p/2}} 2^{-j_\alpha s p} \\
&= \frac{C M^p}{\log_2 n} (n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)} \\
&= o\left((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)}\right).
\end{aligned}$$

For the term I_{322} , using completely analogous argument as in s_{22} in Lemma 4.3 of Li (2003, p45-46), we can show $I_{322} = o((n^{-1} \log_2 n)^{2\alpha/(1+2\alpha)})$. Combining the above terms I_{311} , I_{312} and I_{321} together, this proves the Lemma. \square