Journal of Data Analysis and Information Processing, **2014, 2, 1-5**
Published Online February 2014 (http://www.scirp.org/journal/jdaip)
http://dx.doi.org/10.4236/jdaip.2014.21001

Scientific Research

Estimation of Hazard Function for Censoring Random Variable by Using Wavelet Decomposition and Evaluation of MISE, AMSE with Simulation

Mahmoud Afshari*, Saeed Tahmasebi
Department of Statistics, College of Science, Persian Gulf University, Bushehr, Iran
Email: *Afshari@pgu.ac.ir, Tahmasebi@pgu.ac.ir

Received November 15, 2013; revised December 17, 2013; accepted January 25, 2014

ABSTRACT

Wavelet analysis is one of the mostly new methods of pure and applied mathematics science. In this paper, we use the wavelet method to estimate the hazard function for censoring random variable. We consider the convergence ratio of given estimator. Also we present the simulation in order to test purpose estimator by calculating the mean integrated squared error (MISE) and average mean squared error (AMSE).

KEYWORDS

Wavelet; Estimator; Censoring Random Variable; Mean Square Integral Error; Average Mean Square Error; Simulation

1. Introduction

One of data types, which researchers are extremely interested in, is carrying to the time interval till the occurrence of certain events such as death etc. Every process waiting for a specific event produces survival data. Failure in survival analysis means the occurrence of the event that we were waiting for. The time, which survival is measured after that point, is called the start time.

The failure time which is denoted by T_i, $i=1,2,3,\cdots$, is the time that failure occurs for each individual. It's not always possible to observe the failure time for each individual in such cases that censorship occurs.

Survival function, which is shown by $S(t)$, indicates the ratio of people who survived since the base time which is the point they enter the experiment to the time unit *t* analysis. Hazard function for the failure continuous time is as follows:

$$h_t(t)=\lim_{\Delta t\to 0}\frac{P\left(t\le T\le t+\Delta t\middle| T\ge t\right)}{\Delta t}=\frac{F'(t)}{1-F(t)}=-\frac{S'(t)}{S(t)} \quad (1)$$

*Corresponding author.

In this paper, we obtain estimator hazard function for censoring data by using wavelet method. We evaluate convergence ratio of given estimator by simulation.

2. Estimation of Hazard Function by Using Wavelet Method

Wavelets can be used for transient phenomena analysis or function analysis which sometimes changes rapidly. They are symmetrical and have limited period. A close relationship between wavelet coefficients and some spaces is wavelet bases orthogonally. Also useful properties of them in wavelet issues simplify the computational algorithms. As a result, numerous articles have been published about in statistical science.

The mathematical theorem of wavelets and their application in statistics have been studied as a technique for density function estimator, by Harr [1], Doukhan [2], Antoniadys [3], nonparametric curve estimators by Malat [4], Meyer [5], Daubechies [6], Donoho [7], Kyacharyan and Picard [8], Hall and Patil [9] have found a formula for the Mean Integrated Squared Error of Nonlinear Wavelet based on density estimators. Antoniadys *et al*. [10] achieved the density function estimator and the

hazard function for right-censored data with the wavelets. Daubechies [11] studied and discussed the compactly supported wavelets which produce orthogonal bases. Afshari *et al.* [12-14] studied about density, derivative density function estimator, regression function for the mixing random variables.

Let the nested sequence of closed subspaces; $\mathbf{V}_{j-1} \subset \mathbf{V}_j \subset \mathbf{V}_{j+1} \subset \cdots,\ j \in \mathbf{Z}$, be a multiresolutuon approximation to $\mathbf{L}^2(R)$. Define $\mathbf{W}_j$, $j \in \mathbf{Z}$ to be orthogonal complement of $\mathbf{V}_j$ in $\mathbf{V}_{j+1}$. Wavelets basis for function $f \in \mathbf{L}^2(R)$ as scaling function φ and mother wavelet ψ such that $\{\varphi(x-k)\}_{k\in Z}$ forms an orthogonal basis for $\mathbf{V}_0$ and $\{\psi(x-k)\}_{k\in Z}$ forms an orthonormal basis for $\mathbf{W}_0$. Other wavelets in the basis are then generated by translation of the scaling function and dilations of the mother wavelet by using the relationships:

$$\begin{aligned} \varphi_{m_0,k}(x) &= 2^{m_0/2}\varphi\left(2^{m_0}x-k\right), \\ \psi_{j,k}(x) &= 2^{j/2}\psi\left(2^{j}x-k\right) \end{aligned} \tag{2}$$

Given above Wavelet basis, a function $f \in \mathbf{L}^2(R)$ can be written a formal expansion:

$$f = \sum_{k\in Z}\alpha_{m_0,k}\varphi_{m_0,k} + \sum_{j=m_0}^{\infty}\sum_{k\in Z}\beta_{j,k}\psi_{j,k} \tag{3}$$

where $\alpha_{j_0,k} = \int f(x)\varphi_{j_0,k}(x)\mathrm{d}x, \quad \beta_{j,k} = \int f(x)\psi_{j,k}\mathrm{d}x.$

As for general orthogonal series estimator, Daubechies [4], density estimator can be writhen as:

$$\begin{aligned} \hat{f} &= \sum_{k\in Z}\hat{\alpha}_{m_0,k}\varphi_{m_0,k}(x) + \sum_{j\ge m_0}\sum_{k\in Z}\hat{\beta}_{j,k}\psi_{j,k}(x) \\ &= \mathbf{P}_{m_0} f + \sum_{j\ge m_0}\sum_{k\in Z}\hat{\beta}_{j,k}\psi_{j,k} \end{aligned} \tag{4}$$

where the obvious coefficient estimator can be written:

$$\begin{aligned} \hat{\alpha}_{m_0,k} &= E\left[\varphi_{m_0,k}(X)\right] = \frac{1}{n}\sum_{i=1}^{n}\varphi_{m_0,k}(X_i), \\ \hat{\beta}_{j,k} &= E\left[\psi_{j,k}(X)\right] = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,k}(X_i) \end{aligned} \tag{5}$$

In this article, we divide time axis into two parts, the intervals and the number of events in each interval. We determine number of events and hazard function according to the observations. Then we flatten them separately via linear wavelet density estimation on the whole time and then we calculate the function estimator and evaluate the asymptotic distribution.

Suppose $X_1, X_2, \cdots, X_n$ are failure time of *n* tests that are studied. They are non-negative, independent, identically distributed, with the density function *f* and distribution function *F*. Also suppose that $C_1, C_2, \cdots, C_n$ are corresponding to censored times, non-negative, independent, identically distributed, with the density function g and distribution function G.

Assuming independency of failure times and censored time of the observed random variable, Z_i and the function δ_i and hazard function are shown as below:

$$Z_i = \min(X_i, C_i), \quad \delta_i = I_{(X_i \le C_i)}$$

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

Such that $I_{(A)}$ is indicator function of *A*. For data censoring, if $G(t) < 1$,

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

We assume that,

$$\begin{aligned} L(t) &= P(Z_i \le t) = 1 - P(Z_i > t) \\ &= 1 - P(X_i > t, C_i > t) \\ &= 1 - (1-F(t))(1-G(t)). \end{aligned}$$

Such that $f^*(t) = f(t)(1-G(t))$, then we can write as follows:

$$h(t) = \frac{f^*(t)}{1-L(t)}, \quad L(t) < 1. \tag{6}$$

To estimate $h(t)$ we need the estimator of $f^*(t)$ and $L(t)$.

For estimating $f^*(t)$, we divide the time axis into two parts of small intervals and the amounts of events (0 or 1) in each interval, and then we divide these values to the length of intervals.

Estimation procedures of $f^*(t)$ can be summarized as the following:

Select $\Delta > 0$ and collect the observed failures in $k+1$ intervals with the length Δ and using wavelet estimation on the collected data. We find an estimate of sub density. This means that we calculate the collected wavelet coefficients data on the scale of $j(n)$ by choosing the decomposition level $j(n)$ and then we estimate $f^*(t)$. It is necessary to state the following symbols to show the details:

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

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

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

We figure estimators on the finite interval $[0, \tau]$ in which $\tau < T_L$. Note that if $Z_{(i)}$ is the ordinal order statistic i of the sequence Z_i then, $T_{L_n} = Z_{(n)} \xrightarrow{n\to\infty} T_L$. In fact we suppose $\tau = Z_{(n)}$.

Suppose that *N* is an integer that could be dependent to *n* and the estimated points are as follows:

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

Suppose that $\Delta = \tau 2^{-N}$ and we divide the interval $[0, \tau]$ of time axis to $k+1$ intervals with Δ long $\tau_0 = -\frac{\Delta}{2}, \ \tau_k = t_k - \frac{\Delta}{2}, \ k = 1, \cdots, K, \ \tau_{K+1} = \tau.$

The *k*-th interval is marked by J_k so: $J_k = [\tau_K, \tau_{k+1})$ for $k = 0, \cdots, K-1$, $J_K = [\tau_K, \tau]$.

Now we define the following indicator function that indicates the number of uncensored failures in the time interval J_k: $Y_{ik} = I_{J_k}(Z_i)\delta_i, \quad i = 1, \cdots, n, \quad k = 0, \cdots, K.$ We assume that U_k is the observed failures ratio in the interval J_k, in other words: $U_k = \frac{1}{n}\sum_{i=1}^{n} Y_{ik}, \ k = 0, \cdots, K.$

We smooth the data $\frac{U_k}{\Delta}$ by an appropriate wavelet smoother to find the estimation of f^*.

We can write as the following:

$$\begin{aligned} f^*(t) = & \sum_{k=0}^{2^{j_0}-1} \left\langle f^*, \varphi_{j_0,k} \right\rangle \varphi_{j_0,k}(t) \\ & + \sum_{j \geq j_0} \sum_{\ell=0}^{2^j-1} \left\langle f^*, \psi_{j,\ell} \right\rangle \psi_{j,\ell}(t). \end{aligned} \tag{7}$$

where, $\langle f, g \rangle \equiv \int_{\circ}^{\tau} f(t) g(t) \mathrm{d}t$.

The complex structural polymorphism analysis causes an efficient tree construction algorithm for analysis of functions in V_N with theoretic scale wavelet coefficients $\langle f^*, \varphi_{N,k} \rangle$. However, the integral scale $\langle f^*, \varphi_{N,k} \rangle$ is not well available and we need an initial value for a fast wavelet transform. Antoniadis [4] suggested the following initial amount:

$$\left\langle f^*, \varphi_{N,k} \right\rangle = 2^{-\frac{N}{2}} f^*(t_k) + O\left(2^{-\frac{N}{2}} 2^{-Nm}\right),$$

$$0 \leq k \leq 2^N - 1$$

As a result a reasonable estimate for image of f^* with clarity *N* is: $\tilde{f}_N^*(t) = 2^{-\frac{N}{2}} \sum_{k=0}^{K} \frac{U_k}{\Delta} \varphi_{N,k}(t) \cdot$

If we assume that the collected values U_k which are equal to the estimators of $\tilde{f}_N^*$, are in Sobolev space $W^m([0,\tau])$ and φ is regular of degree m. We estimate the unknown function f^* as follows to level the data with a better rate for the sample size n and the sequence $j(n) \leq N$:

$$\hat{f}_n = P_{V_{j(n)}}^{\tilde{f}_N^*}. \tag{8}$$

That it is the orthogonal image of $\tilde{f}_N^*$ on the leveler approximation space $V_{j(n)}$.

Now we consider an appropriate consistent estimator of $L(t)$, and finally we estimate the Hazard function.

We assume that $Z_1, Z_2, Z_3, \cdots, Z_n$ has distribution function $L(t)$ and density function $l(t)$.

For estimating of $L(t)$, we use an empirical distribution $L_n(t)$ as the following:

$L_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{(Z_i \leq t)}$. $\hat{L}_n(t) = \int_0^t \hat{l}_n(x)\mathrm{d}x, \ t \in [0,\tau]$ Such that $\hat{l}_n(t)$ is Histogram estimator of $l(t)$. Suppose that, $\varphi(t) = \mathrm{I}_{[0,\tau]}(t)$, we can write:

$$\begin{aligned} & \hat{l}_n(t) \\ & = \frac{1}{n}\sum_{i=1}^{n} 2^{\frac{\tilde{j}(n)}{2}} \varphi_{\tilde{j}(n),0}(t - z_i) \\ & = \frac{1}{n}\sum_{i=1}^{n} 2^{\frac{\tilde{j}(n)}{2}} 2^{\frac{\tilde{j}(n)}{2}} \varphi\left(2^{\tilde{j}(n)}(t - z_i)\right) \\ & = \frac{1}{n}\sum_{i=1}^{n} 2^{\tilde{j}(n)} \varphi\left(2^{\tilde{j}(n)}\right) \varphi\left(2^{\tilde{j}(n)}(t - z_i)\right) \end{aligned}$$

Suppose that $\tilde{j}(n) \to \infty$ as $n \to \infty$, then we define:

$\Phi_{\tilde{j}(n),k}(t) = 2^{\frac{j}{2}} \int_0^t \varphi_{j,k}(x)\,\mathrm{d}x = 2^j \int_{\circ}^{t} \varphi\left(2^j x - k\right)\mathrm{d}x,$ so we can write as the following:

$$\begin{aligned} & \hat{L}_n(t) \\ & = \frac{1}{n}\sum_{i=1}^{n} \Phi_{\tilde{j}(n),0}(t - z_i) \\ & = \frac{1}{n}\sum_{i=1}^{n} 2^{\frac{\tilde{j}(n)}{2}} \int_0^{t-z_i} \varphi_{\tilde{j}(n),0}(x)\mathrm{d}x \\ & = \frac{1}{n}\sum_{i=1}^{n} 2^{\frac{\tilde{j}(n)}{2}} \int_0^{t-z_i} 2^{\frac{\tilde{j}(n)}{2}} \varphi\left(2^{\tilde{j}(n)}(x)\right)\mathrm{d}x \\ & = \frac{1}{n}\sum_{i=1}^{n} 2^{\tilde{j}(n)} \int_0^{t-z_i} \varphi\left(2^{\tilde{j}(n)}(x)\right)\mathrm{d}x. \end{aligned} \tag{9}$$

By substituting Equation (9) in Equation (8), we obtain the estimator $\hat{h}_n(t) = \frac{\hat{f}_n(t)}{1 - \hat{L}_n(t)}$.

Theorem: Suppose that the sub density $h(t)$ is a continuous function on $[0,\tau]$ and it's m times differentiable, If $j(n) = \frac{1}{2m+1}\log_2(n)$ and $N \geq \frac{m}{2m+1}\log_2(n)$ then,

$$MISE\left(\hat{h}_n(t)\right) = E\left[\left(\hat{h}_n(t) - h(t)\right)^2\right] = \frac{\hat{f}_n(t) - f^*(t)}{1 - L(t)},$$

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

as $n \to \infty$

Proof:

$$\hat{h}_n(t)=\frac{E\left(\hat{f}_n(t)\right)+\left\{\hat{f}_n(t)-E\left(\hat{f}_n(t)\right)\right\}}{1-E\left(\hat{L}_n(t)\right)+E\left(\hat{L}_n(t)\right)-\hat{L}_n(t)}$$
$$=\frac{E\left(\hat{f}_n(t)\right)+\left\{\hat{f}_n(t)-E\left(\hat{f}_n(t)\right)\right\}}{E\left(1-\hat{L}_n(t)\right)}$$
$$+\frac{E\left(1-\hat{L}_n(t)\right)}{1-E\left(\hat{L}_n(t)\right)+E\left(\hat{L}_n(t)\right)-\hat{L}_n(t)}$$
$$=\frac{E\left(\hat{f}_n(t)\right)+\left\{\hat{f}_n(t)-E\left(\hat{f}_n(t)\right)\right\}}{E\left(1-\hat{L}_n(t)\right)}$$
$$+\left(\frac{1-E\left(\hat{L}_n(t)\right)+E\left(\hat{L}_n(t)\right)-\hat{L}_n(t)}{E\left(1-\hat{L}_n(t)\right)}\right)^{-1}$$
$$=\left\{\frac{E\left(\hat{f}_n(t)\right)}{E\left(1-\hat{L}_n(t)\right)}+\frac{\left\{\hat{f}_n(t)-E\left(\hat{f}_n(t)\right)\right\}}{E\left(1-\hat{L}_n(t)\right)}\right\}$$
$$+\left(1+\frac{E\left(\hat{L}_n(t)\right)-\hat{L}_n(t)}{E\left(1-\hat{L}_n(t)\right)}\right)^{-1}$$

By using Chung-Smirnov property and Taylor's theorem we can write as the following:

$$\hat{h}_n(t)=\frac{E\left(\hat{f}_n(t)\right)}{E\left(1-\hat{L}_n(t)\right)}+\frac{\left\{\hat{f}_n(t)-E\left(\hat{f}_n(t)\right)\right\}}{E\left(1-\hat{L}_n(t)\right)}$$
$$-\frac{\left\{E\left(\hat{L}_n(t)\right)-\hat{L}_n(t)\right\}E\left(\hat{f}_n(t)\right)}{\left(E\left(1-\hat{L}_n(t)\right)\right)^2}$$
$$+O\left[\left|\hat{f}_n(t)-E\left(\hat{f}_n(t)\right)\right|+\left|E\left(\hat{L}_n(t)\right)-\hat{L}_n(t)\right|\right]. \quad (10)$$

$$\frac{E\left(\hat{f}_n(t)\right)}{E\left(1-\hat{L}_n(t)\right)^{-1}}=\frac{\left(f^*(t)+\left[E\left(\hat{f}_n(t)\right)-f^*(t)\right]\right)}{1-L(t)}$$
$$\times\left\{1-\frac{E\left(\hat{L}_n(t)\right)-L(t)}{1-L(t)}+O\left(\left|E\left(\hat{L}_n(t)\right)-L(t)\right|\right)\right\} \quad (11)$$

By using Equations (10) and (11), we can write:

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

then the proof is completed.

3. Numerical Computation and Simulation

In this section, we simulate $\hat{f}_n(t_k)$ and $\hat{h}_n(t_k)$ on the data of size n by using Semlayt's wavelet. We consider convergence ratio of given estimator by computing of average mean square error of given estimators. We use R software and wavelet package for simulations.

Example 1: We generate $X_1, X_2, X_3, \cdots, X_n \sim \Gamma(5,1)$ and $C_1, C_2, C_3, \cdots, C_n \sim E(6)$ from the samples of size $n=400$ and $n=600$ with K = 16, $K=32$, $K=64$ and $\Delta=0.05$ for optimal surface $j=2$.

The solid line in the **Figure 1** displays the wavelet estimate of hazard function with the denoted line representing the true hazard rate

The results in **Table 1** display the average mean square errors of hazard function estimator for sample sizes $n=400$ and $n=600$.

Example 2: Suppose $X_1, \cdots, X_n \sim f = 0.6Y + 0.4W$, where $Y \sim LN(0,1)$ and $W \sim N(3,0.04)$. We generate $C_1, C_2, C_3, \cdots, C_n \sim E(3)$ from sample size of n = 400 and n = 600 with K = 16, K = 32, K = 64 and $\Delta=0.05$.

The solid line in the **Figure 2** displays the wavelet estimate of hazard function with the denoted line representing the true hazard rate.

The results in **Table 2** display the average mean square

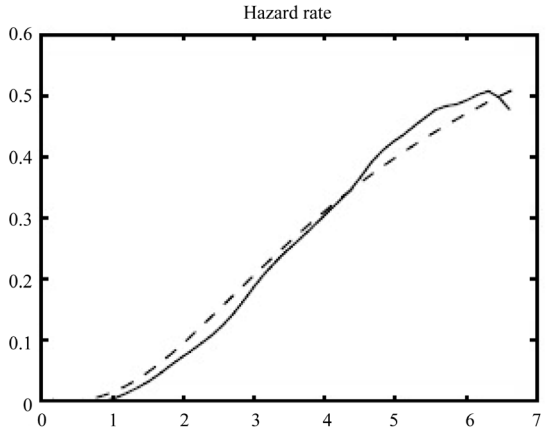

Figure 1. The panel in Figure 1 displays the wavelet estimator of hazard function $\hat{h}_n(t_k)$ with the denoted line representing the true hazard rate.

Table 1. Average mean square errors of hazard function estimator.

$$AMSE(h)=\frac{1}{K}\sum_{k=0}^{K}\left(\hat{h}_n(t_k)-h(t_k)\right)^2$$

k	n = 400	n = 600
16	65.6	56.3
32	79.2	55.9
64	114.3	98.9

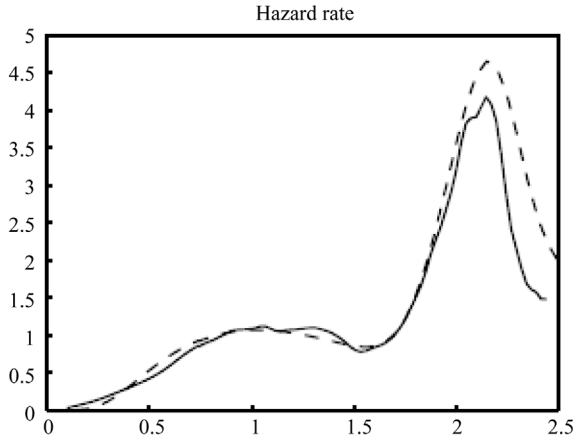

Figure 2. The solid line in the panel displays the wavelet estimate of hazard function with the denoted line representing the true hazard rate.

Table 2. Average mean square errors of hazard function estimator.

	$AMSE(h)=\frac{1}{K}\sum_{k=0}^{K}\left(\hat{h}_n(t_k)-h(t_k)\right)^2$	
K	n = 400	n = 600
16	3065	3100
32	4092	1875
64	2097	1985

errors of hazard function estimator for sample sizes n = 400 and n = 600.

Acknowledgements

The support of Research Committee of Persian Gulf University is greatly acknowledged.

REFERENCES

[1] A. Haar, "Zur Thorie der Orthogonal Functioned-System," *Annals of Mathematics*, Vol. 69, No. 3, 1910, pp. 331-371. http://dx.doi.org/10.1007/BF01456326

[2] P. Doukhan, "Mixing Properties and Examples," Springer-Verlag, New York, 1995.

[3] A. Antoniadis, "Smoothing Noisy Data with Tapered Coiflet Series," *Scandinavian Journal of Statistics*, Vol. 23, 1996, pp. 313-330.

[4] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7, 1989, pp. 674-693. http://dx.doi.org/10.1109/34.192463

[5] Y. Meyer, "Ondelettes et Operateurs," Hermann, Paris, 1990.

[6] I. Daubechies, "Ten Lectures on Wavelets," SIAM, Philadelphia, 1992.

[7] D. L. Donoha and I. M. Johnstone, "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika Journal*, Vol. 81, No. 3, 1994, pp. 425-455. http://dx.doi.org/10.1093/biomet/81.3.425

[8] G. Kerkyacharian and D. Picard, "Density Estimation by Kernel," *Probability and Letters*, Vol. 18, No. 4, 1993, pp. 327-336. http://dx.doi.org/10.1016/0167-7152(93)90024-D

[9] P. Hall and P. Patil, "Formula for Mean Integrated Squared Error of Non-Linear Wavelet Based Density Estimators," *Annals of Statistics*, Vol. 23, No. 3, pp. 905-928. http://dx.doi.org/10.1214/aos/1176324628

[10] A. Antoniadis, G. Gregoire and G. P. Nason, "Density and Harzard Rate Estimation for Right Censored Data Using Wavelet Methods," *Journal of Royal Statistical Society, Series B*, Vol. 61, No. 1, 1999, pp. 63-84. http://dx.doi.org/10.1111/1467-9868.00163

[11] I. Daubechies, "Orthogonal Bases of Compactly Supported Wavelets," *Communication in Pure and Applied Mathematics*, Vol. 41, No. 7, 1988, pp. 909-996. http://dx.doi.org/10.1002/cpa.3160410705

[12] M. Afshari, "A Fast Wavelet Algorithm for Analyzing of Signal Processing and Empirical Distribution of Wavelet Coefficients with Numerical Example and Simulation," *Communication of Statistics-Theory and Methods*, Vol. 42, No. 22, 2013, pp. 4156-4169. http://dx.doi.org/10.1080/03610926.2011.642917

[13] M. Afshari, "Wavelet Density Estimation of Censoring Data and Evaluate of Mean Integral Square Error with Convergence Ratio and Empirical Distribution of Given Estimator," 2013, under print.

[14] H. Doosti, M. Afshari and H. A. Niroomand, "Wavelets for Nonparametric Stochastic Regression with Mixing Stochastic Process," *Communication of Statistics-Theory and Methods*, Vol. 37, No. 3, 2008, pp. 373-385. http://dx.doi.org/10.1080/03610920701653003