Journal of Engineering and Applied Sciences 14 (1): 242-246, 2019
ISSN: 1816-949X

Density Estimation for Right Censored Data Using Hybrid Breslow and Semi-Symmetric Wavelet

[1]Ali Talib Mohammed and [2]Iden Hussein
[1]Department of Mathematics, College of Education For Pure Sciences (Ibn AL-Haitham),
[2]Department of Mathematics, College of Sciences for Women, University of Baghdad, Baghdad, Iraq

Abstract: This study introduce a wavelet method estimator of the density function based on the projection property of the father wavelet functions $\{\Phi_{J,k}(x)\}$, $0\leq k\leq 2^J-1$, $J\geq 0$ on the subspace $V_J\subseteq L^2$ for non-parametric randomly censoring data with assuming that the density function belong to $L^2(R)$ has no specific parametric distribution. The technique using coiflet (coifN, N = 1, 2,..., 9) wavelet which is semi-symmetric orthogonal and Breslow estimator for finding the cumulative function. To compare and determine the best results using the Mean Square Error (MSE) for estimating the density function of real right censoring data of 214 patients called home nursing data.

Key words: Breslow estimator, censored data, cumulative functions, density estimation, multiresolution analysis, wavelet methods

INTRODUCTION

One of the major challenges faced by researchers in statistical and other studies is to deal with non-parametric data. The important objective is how to estimate a function of these data without any information about this function. In statistical studies, non-parametric data are one of these challenges in terms of estimating density function. Most of the studies that investigated this aspect were based on the analysis and study of data properties. An significant property is that $X_1, X_2, ..., X_n$ for n samples are to be independent. kernel and nearest neighbors technics are the most popular methods to non-parametric density function. Wavelet series could be considered as good approach for function estimator, since, the utilize of wavelet in solving this problem gives an important advantage because it build up functions belong to $L^2(R)$. A first used of wavelets in statistics were introduced by a collection of articles such as Donoho *et al*. (1996) and Van Fleet (2011). One of the advantages of wavelets is that have significant and distinct characteristics for time and frequency which makes them an important tool for estimating the functions Daubechies (1992) and Van Fleet (2011). Antoniadis *et al*. (1999) presented a wavelet-based method for estimating hazard rate and density function for right censoring survival data. Chaubey *et al*. (2010) estimationfor derivative of density function using wavelet-based for randomly censoring data. Cai and Liang (2015) estimated the density function used non-linear wavelet method for of truncated and dependent observing data. Chesneau and Willer (2015), estimated the cumulative function for non-parametric data and construct a new adaptive estimator based on a warped wavelet basis and a hard thresholding rule. Abbaszadeh *et al*. (2013), estimated the density function and it's derivatives for a sample of multiplicatively censored random variables by projection linear wavelet estimator and non-linear term-by-term selection wavelet. Afshari (2014) have done some researches about density function estimator use wavelet method for estimating the density function for censoring data, and evaluated the mean integrated squared error. Chesneau and Doosti (2016) developed a new estimator g (x, m) based on wavelet methods of multivariate discrete and continuous density function. Grez and Vidakovic (2018), estimated the density function using empirical approach linear estimator based on an orthogonal projection wavelet with Kaplan-Meier estimator of randomly censored data and proposed the multiresolution space index $J = \log_2 N - \log_2(\log(N))$. In clinical trials or survival studies, patients who need to follow up for different periods of time may range from a few days to several years in addition to the main event being death or survival. In randomly censored medical data patients inter the study in different times during the period, so, their exact survival times are known. Besides that,

Corresponding Author: Ali Talib Mohammed, Department of Mathematics,
College of Education For Pure Sciences (Ibn AL-Haitham), University of Baghdad, Baghdad, Iraq

censored time is specify for each individual. Before the end of study some patient may withdraw and lost to follow up.

Let, X_1, X_2, X_3, ..., X_n be an independent and identically distributed (i.i.d.) survival times with unknown density function f. let C_1, C_2, ..., C_n be i.i.d. censoring times unknown density function g. it is presumed that for i = 1, 2, …, n X_i and C_i are typically statistical independence. The observing is for $\{Z_i, \delta_i\}$ which is an i.i.d. sequences, such that $Z_i = \min \{X_i, C_i\}$ where:

$$\delta_i = \begin{cases} 1 \text{ if } X_i \le C_i \\ 0 \text{ if } X_i > C_i \end{cases}$$

This study, present a linear wavelet estimator method dependent on semi-symmetric orthogonal coiflte wavelet which is the main part and the primary part is bersilow estimator for cumulative function. One of the advance of this method is that used of the orthogonal projection on the multiresolution space V_J with index $J = \log_2(n/\log_{10}(n))$ to estimator the density function for 214 patients of home nursing data which was presented by Morris *et al.* Moreover, using the Mean Square Error (MSE) to collect the average error between the original values and estimated values and choosing the minimum square error.

Wavelet: Wavelets are considering as a new class of functions that are well localized in time and frequency. Moreover, the wavelet is rapidly decaying wave like oscillation that has zero mean and it exists for the finite duration. Us a transformations wavelets could be used in two types of Discrete Wavelet Transformation (DWT) and Continuous Wavelet Transformation (CWT). The method that will be displayed focus on (DWT). Approximation and estimation of functions is one of the important and good uses of wavelets.

A multiresolution analysis $L^2(R) = \{f: R \to R, \int_{-\infty}^{\infty} |f(x)|^2 dx < \infty\}$ contains of subspaces $\{V_J\}_{J\in Z}$ with $\overline{\cup_{J\in Z} V_J} = L^2(R)$ and $\cap_{J\in Z} V_J = 0$. The different subspace $W_J = V_{J+1} \ominus V_J$ for all $J \in Z$ is a subspace of $L^2(R)$.

The sequence of functions $\{\Phi_{J,k}(x)\}$ and $\{\Psi_{J,k}(x)\}$, $0 \le k \le 2^J-1$, $J \ge 0$ are two basis for the subspaces V_J and W_J respectively where $\Phi_{j,k}(x) = 2^{J/2}\, \Phi(x-k)$ and $\Psi_{J,k}(x) = 2^{1/2}\Psi(x-k)$, moreover $\Phi_{J,k}(x)$ and $\Psi_{J,k}(x)$ are called father and mother wavelet, respectively.

Preliminary: Let $\{X_i, i = 1, 2, ..., n\}$ be i.i.d of non-negative random function with density function (pdf) f(.) and g(.), cumulative functions (cdf) F(.) and G(.), respectively. Let $Z_i = \min \{X_i, C_i\}$ be the survival times (observed times)with the indicator function $\delta_i = 1_{Xi \le ci}$ and 0 otherwise, so there is no censoring for ith observed time if $\delta_i = 1$.

Now, assumed that $\beta = \max \{Z_i, i = 1, 2, ..., n\}$ and to make sure that all observed times Z_i belong to [0, 1], put $\hat{Z}_i = 1/\beta Z_i$ and $\{\hat{Z}_{(i)}, \delta_{(i)}\}$ be the ranked of $\{\hat{Z}_i, \delta_i\}$.

Based on hierarchical of multiresolution, any function f belong to $L_2(R)$ could be written as follows:

$$f(x) = \sum_k W_\Phi(j_0, k)\Phi_{j0,k}(x) + \sum_{j=j0}^{\infty}\sum_k W_\psi(j, k)\psi_{j,k}(x) \quad (1)$$

where, j = 1, 2, …, kϵZ, j_0 is an arbitrary starting scale and j0$\le$j. Because of that f(.) is a probability density function, the coefficients $\Phi(j_0, k)$ and $W_\Psi(j, k)$ can be expressed as:

$$W_\Phi(j_0, k) = E\left[\Phi_{j0,k}(X)\right] \quad (2)$$

$$W_\psi(j, k) = E\left[\psi_{j,k}(X)\right] \quad (3)$$

And as mention above, since, f(.) is unknown, then: $\hat{W}_\Phi(j_0,k) = 1/n\sum_x f(x)\Phi_{j0,k}(x)$ is called "Approximation" coefficients.

$\hat{W}_\Psi(j,k) = 1/n\sum_x f(x)\Psi_{j,k}(x)$ is called "Detail" coefficients. Ψ and Φ are known as Mother Wavelet and Father Wavelet, respectively.

From Eq. 1 can see that j is start j_0 and end with infinity. Based on that, f(x) could be approximated from j_0-$\hat{J}$. The value of scale index $\hat{J} = \log_2(n/\log_{10}(n))$ and k is belong to the interval $[0,2^{\hat{J}}-1]$. Therefore, Eq. 1 reformulate as follows:

$$f(x) = \left(\frac{1}{n}\sum_x f(x)\Phi_{j0,k}(x)\right)\Phi_{j0,k}(x) + \sum_{j=j0}^{J}\sum_k\left(\frac{1}{n}\sum_x f(x)\psi_{j,k}(x)\right)\psi_{j,k}(x) \quad (4)$$

Estimation of $f_j(.)$ for the observed times: The interest here about the work that was introduced by German *et al.* respectively, they estimated the density function based on a hybrid between Kaplan-Meier estimator and Symmetric (Daubechies) wavelet. So, far this study focus to estimator the density function for observed times $\{\hat{Z}_{(i)}, \delta_{(i)}\}$. Firstly, assume that G(Z)ϵ(0, 1) for all Z and G go to infinity for non-censored data. Consequently, this will lead to $f_z = f(x)$, moreover, $W\Phi = 1/n\sum_{i=1}^{n}\Phi_{j,k}(x)$ and $W\Psi = 1/n\sum_{i=1}^{n}\Psi_{j,k}(x)$. The estimation of density function $f_j(x)$ could be written as:

$$f_j(x) = \sum_{k=2}^{2\hat{J}-1}\widehat{W}_\Phi\Phi_{j0,k}(x) + \sum_{j=j0}^{\hat{J}}\sum_{k=0}^{2\hat{J}-1}\widehat{W}_\Psi\Psi_{j,k}(x) \quad (5)$$

Equation 5 gives us an ability to find the density function from random variable X of fully observed lifetimes.

MATERIALS AND METHODS

Methodology of estimation based on orthogonal projection: One of the good advantage for wavelet is to estimate any function $f \epsilon L_2(R)$ which approximate function based on orthogonal projection. Basically, for fixed scale $\hat{J}$ the orthogonal projection of $f_j(x)$ onto the subspace V_j is denoted $p(f_j(x))$ and given as:

$$p(f_{\hat{j}}(x)) = \sum_{k=0}^{2^{\hat{j}}-1} \langle f(x), \Phi_{j0,k}(x) \rangle \Phi_{j0,k}(x) \quad (6)$$

Denoting the orthogonal projection density function using $\hat{f}_{\hat{j}}(x)$. Based on Eq. 6, need to find the coefficient $\langle f(x), \Phi_{j0,k}(x) \rangle$, so, let first denoted it as $\varsigma\Phi$. Moreover, since, f(.) is unknown density function, for that use the Cdf's F and G to collect $\varsigma\Phi$ from the observed data $\{Z_i, \delta_i\} i = 1, 2, ..., n$ the joint distribution of (Z, δ) is:

$$P(Z \le z, \delta = 1) = \int_{-\infty}^{z} (1-G(x)f(x)dx \quad (7)$$

$$P(Z \le z, \delta = 0) = \int_{-\infty}^{z} (1-G(x)f(x)dx + \int_{z}^{\infty} G(z)f(x)dx \quad (8)$$

Dependent on Eq. 7 and 8:

$$f_z(Z) = f_x(z)(1-G_c(z)) + g_c(z)(1-F_x(z)) \quad (9)$$

As a result for Eq. 9:

$$f_x(z) = \frac{fZ(z)}{1-G_C(z)} - \frac{g_C(z)(1-F_x(z))}{1-G_C^{(z)}} \quad (10)$$

From Eq. 10, it possible to express and formed $\varsigma_\Phi = \langle f(x), \Phi_{j0,k}(x) \rangle$ as:

$$\varsigma_\Phi = \int_0^1 \frac{fz(Z)}{1-G_C(z)} - \frac{g_C(z)(1-F_x(z))}{1-G_C(z)} \Phi_{j0,k}(x)d(x)$$
$$\varsigma_\Phi = E\left[\frac{\Phi_{j0,k}(Z)}{1-G(Z)}\right] - E\left[\frac{1-F(Z)\Phi_{j0,k}(Z)}{1-G(Z)}\right] \quad (11)$$

Using the approach $\widehat{W}_\Phi(j_0,k) = 1/n \sum_x f(x)\Phi_{j0,k}(x)$ and $\widehat{W}_\Psi(j,k) = 1/n \sum xf(x)\Psi_{j,k}(x)$ for $0 \le G(Z_i), i = 1, 2, ..., n$:

$$\varsigma_\Phi = n\text{-}1\sum_{i=1}^{n} \frac{\Phi_{j0,k}(Z_i)}{1-G(Z_i)} - n\text{-}1\sum_{i=1}^{n} \frac{1_{\delta i=0}(1-F(Z_i))\Phi_{j0,k}(Z_i)}{1-G(Z_i)} \quad (12)$$

Now, $\hat{F}(\hat{Z}_{(i)})$ and $G(\hat{Z}_{(i)})$ for i = 1, 2, .., n can be estimated using Breslow estimator for survival function as follows:

$$\hat{F}(\hat{Z}_{(i)}) = \sum_{r=1}^{i} \frac{1-\delta_{(r)}}{n-r+1} e^{-\left(\sum_{s=1}^{r-1} \frac{1-\delta(s)}{n-s+1}\right)} \quad (13)$$

$$\hat{G}(\hat{Z}_{(i)}) = \sum_{r=1}^{i} \frac{1-\delta_{(r)}}{n-r+1} e^{-\left(\sum_{s=1}^{r-1} \frac{1-\delta(s)}{n-s+1}\right)} \quad (14)$$

$$\xi_i = \frac{1}{1-\hat{G}(\hat{Z}_{(i)})} - \frac{I_{\delta i=0}\left(1-\hat{F}(\hat{Z}_{(i)})\right)}{1-\hat{G}(\hat{Z}_{(i)})} \quad (15)$$

Rewrite Eq. 12 as follows:

$$\varsigma_\Phi = \frac{1}{n}\sum_{i=1}^{n} \xi_i \Phi_{\hat{j}k}(\hat{Z}_{(i)}) \quad (16)$$

Finally, the estimate of density function $\tilde{f}_{\hat{J}}(x)$ for chosen scale index $\hat{J} = \log_2(n/\log_{10}(n))$ and orthogonal semi-symmetric coiflet wavelet coifN (N = 1, 2, …, 9) can be formed as:

$$\tilde{f}_{\hat{J}}(x) = \sum_{k=2}^{2^{\hat{J}}-1} \varsigma_\Phi \Phi_{\hat{J},k}\left(\hat{Z}_{(i)}\right) \quad (17)$$

Generally, applying this method for any types of right censored data with considering the density function (g) for the censoring indicator in another word for i = 1, 2, ..., n:

$$C_i = \begin{cases} 1 \text{ if censoring} \\ 0 \text{ if uncensoring} \end{cases}$$

Therefore, the observing times take the form $\{X_i, \delta_i\}$

Data application: For data application, the data named is nursing home data which was first introduced by Morris, Norton and Zhou. Data were collected for patients in a nursing home for the elderly between 1980-1982. The original study contains 1601 patients of home nursing and collected by the National Center for Health Services. For application using a subset of original data (n = 214).

The method of data processing for estimating the probability density function method in this study depends on two parts. The first involves the use of a semi-symmetric orthogonal coiflet wavelet (coflN, N = 1, 2, …, 9) which is considered the main part as it helps in finding the father wavelet $\{\Phi_{J,k}(x)\}$, $0 \le k \le$ which will be the main element in Eq. 16. The second part which includes the finding of the coefficient (ς_Φ) in Eq. 15, depends on the father wavelet as well as the Breslow estimator for surviving function. Furthermore, considering the Mean Square Error (MSE) to find the average squared difference between the estimated values and original values (Table 1).

Table 1: Nursing home data of 214 patient

X	δ	X	δ	X	δ	X	δ	X	δ	X	δ
89	1	540	0	25	1	211	1	44	1	37	1
27	1	222	1	73	1	100	1	480	0	3	1
156	1	400	1	270	1	661	0	57	1	84	1
12	1	471	0	33	1	19	1	44	1	148	1
40	1	204	1	96	1	69	1	9	1	365	0
1	1	60	1	2	1	32	1	260	1	18	1
5	1	407	0	211	1	197	1	6	1	168	1
122	1	139	1	25	1	49	1	172	1	18	1
306	1	48	1	487	0	470	1	41	1	281	1
21	1	507	0	9	1	572	0	298	1	24	1
597	1	250	1	46	1	82	1	176	1	18	1
140	1	54	1	66	1	156	1	370	1	64	1
47	1	611	0	310	1	537	0	1	1	90	1
635	0	501	0	721	0	1	1	0	1	133	1
724	0	462	0	185	1	54	1	180	1	547	1
707	0	717	0	129	1	2	1	494	0	37	1
696	0	519	1	617	0	277	1	261	1	234	1
435	1	24	1	152	1	25	1	44	1	149	1
47	1	69	1	30	1	474	0	465	0	39	1
607	0	82	1	134	1	45	1	470	0	13	1
598	0	10	1	113	1	28	1	14	1	29	1
576	0	85	1	444	0	30	1	375	1	224	1
141	1	463	0	722	0	457	0	710	0	396	0
537	0	379	1	709	0	198	1	483	1	121	1
1	1	342	1	642	0	72	1	161	1		
480	0	365	0	494	1	114	1	149	1		
50	1	182	1	306	1	15	1	285	1		
20	1	131	1	120	1	409	0	613	0		
123	1	10	1	463	0	27	1	6	1		
15	1	19	1	91	1	8	1	428	1		
0	1	133	1	421	0	258	1	592	0		
43	1	120	1	155	1	22	1	584	0		
47	1	209	1	725	0	96	1	180	1		
165	1	151	1	191	1	382	0	44	1		
696	0	4	1	25	1	673	0	24	1		
38	1	471	0	687	0	624	0	506	0		
11	1	303	1	1	1	36	1	487	0		
585	0	44	1	164	1	2	1	7	1		

RESULTS AND DISCUSSION

The final results of the general content are good results dealing here with unbiased data for any distribution (non-parametric data). In general, the percentages of results can be considered good. In addition to the mean square error give as a good result. From Table 1 observes, there are negative percentages of $\tilde{f}_j(x)$ to explain that, the results come from a non-preapprehension dataset which is Not commensurate with the wavelet type used in the estimation and the expansion of the periodic process. Furthermore, the problem with the use of the wavelet in estimating the probability density function. It is concluded that not all the results are positive and are not concentrated between 0 and 1. From Table 2 observes, there are negative percentages of $\tilde{f}_j(x)$ to explain that, the results come from a non-preapprehension dataset which is not commensurate with the wavelet type used in the estimation and the expansion of the periodic process. Furthermore, the problem with the use of the wavelet in estimating the probability density function. It is concluded that not all the results are positive and are not concentrated between 0 and 1. These problems are due to the tails of the probability density function taking into consideration that these values come from areas of weakens representation, an additional reason is use of parent wavelet functions (Antoniadis, 1997). For MSE results, Coif 9 give as the best result, however, it's come with the high percentage for negative estimator results. Furthermore, the best percentage of negative estimator values it's come of using coif 3 with MSE = 0.5804.

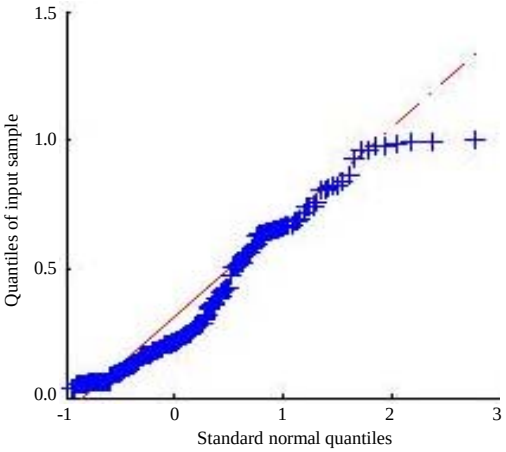

Fig. 1: QQ_plot for density estimator for standard normal of n = 214

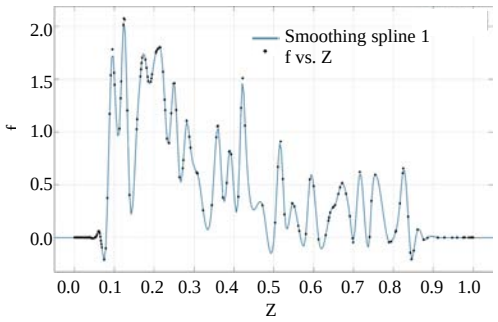

Fig. 2: The smoothing spline1 of $\{(Z_i, \tilde{f}_j) i = 1, 2, ...,\}$

Figure 1 showed the QQ_plot or probability plotthe density estimator for home nursing data of 214 patient compared with standard normal using coif3 wavelet, the estimates are consistent with a normally distributed random variable.

To show more information about the results, Table 2 show the using of smoothing spline (cftool.mat) for the pair $z_i, \tilde{f}_j$ to bridge the gap between these nods.

Table 2: Fitting types of smoothing spline of four parameters

Fit type (RMSE)	Smoothing parameter	----------SSE----------		R^2	Adj. R^2
Smoothing spline 10.01107	1	0.01292		0.9998	0.99970
Smoothing spline 2	0.99999999	0.02909	0.9997	0.9994	0.01606
Smoothing spline 30.06187	0.99999995		0.5397	0.9937	0.99050
Smoothing spline 4	0.99999992	0.9665	0.9888	0.9838	0.08094

CONCLUSION

This study has provided a method to estimate the probability density function for right censoring data, this method is dependent on the construction of two parts, the first one which can be considered as the main part where it is used a semi_symmetric orthogonal wavelet (Coiflet wavelet). While the second part is the use of the Breslow estimator in the estimation function to find G(Z). The probability density function was expressed using the projection property of the father wavelets $\{\Phi_{J,k}(x)\}$, $0 \le k \le 2^J-1$, $J \ge 0$ on the subspace V_J, depending on the correct selection J. Real data was applied of home nursing data for 214 patient to estimate the density function, besides that to collect the error, collect the Mean Square Error (MSE). Moreover, the using of Coiflet wavelet comes with using of all types (coifN, n = 1, 2, …, 9) except (coif6) to estimate the density function which assumed it belong to the multire solution subspace V_J with index $J = \log_2(n/\log_{10}(n))$. The best results got from using a wavelet type coif3 with MSE = 0.5804 and with the least percentage of negative results, also In this study, and through the proposed method for estimating the density function, negative values were observed for some estimates. In addition, the existence of other estimates is not integrated into 1. Finally, MATLAB (Ra2012) was used for programming.

REFERENCES

Abbaszadeh, M., C. Chesneau and H. Doosti, 2013. Multiplicative censoring: Estimation of a density and its derivatives under the Lp-risk. Revstat, 11: 255-276.

Afshari, M., 2014. Wavelet density estimation of censoring data and evaluate of mean integral square error with convergence ratio and empirical distribution of given estimator. Applied Math., 5: 2062-2072.

Antoniadis, A., 1997. Wavelets in statistics: A review. J. Italian Statist. Soc., 6: 97-130.

Antoniadis, A., G. Gregoire and G. Nason, 1999. Density and hazard rate estimation for right-censored data by using wavelet methods. J. Royal Stat. Soc., 61: 63-84.

Cai, J.J. and H.Y. Liang, 2011. Nonlinear wavelet density estimation for truncated and dependent observations. Intl. J. Wavelets Multiresolution Inf. Process., 9: 587-609.

Chaubey, Y.P., H. Doosti, E. Shirazi and R.B. Prakasa, 2010. Linear wavelet-based estimation for derivative of a density under random censorship. J. Iran. Stat. Soc., 9: 41-51.

Chesneau, C. and H. Doosti, 2016. A note on the adaptive estimation of a conditional continuous-discrete multivariate density by wavelet methods. Chin. J. Math., 2016: 1-16.

Chesneau, C. and T. Willer, 2015. Estimation of a cumulative distribution function under interval censoring case 1 via warped wavelets. Commun. Stat. Theor. Methods, 44: 3680-3702.

Daubechies, I., 1992. Ten Lectures on Wavelets: CBMS-NSF Regional Conference Series in Applied Mathematics. Vol. 33, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, ISBN-13:978-0-898712-74-2, Pages: 355.

Donoho, D.L., I.M. Johnstone, G. Kerkyacharian and D. Picard, 1996. Density estimation by wavelet thresholding. Ann. Stat., 24: 508-539.

Grez, G.A.S. and B. Vidakovic, 2017. An empirical approach to survival density estimation for randomly-censored data using Wavelets. J. Theor. Appl. Stat., 1: 1-31.

Grez, G.A.S. and B. Vidakovic, 2018. Empirical wavelet-based estimation for non-linear additive regression models. J. Am. Stat. Assoc., 1: 1-39.

Van Fleet, P.J., 2011. Discrete Wavelet Transformations: an Elementary Approach with Applications. John Wiley & Sons, Hoboken, New Jersey, USA., ISBN:9781118030660, Pages: 564.