Cihan University-Erbil Scientific Journal (CUESJ)

Research Article

A Comparison Between New Modification of Aadaptive Nadaraya-Watson Kernel and Classical Aadaptive Nadaraya-Watson Kernel Methods in Nonparametric Regression: A Simulation Study

Hazhar T. A. Blbas[1], Wasfi T. Kahwachi[2]

[1]Department of Statistics, College of Administration and Economics, Salahaddin University-Erbil, Kurdistan Region, Iraq, [2]Research Center, Tishk International University-Erbil, Kurdistan Region, Iraq

ABSTRACT

Nonparametric kernel estimators are mostly used in a variety of statistical research fields. Nadaraya-Watson kernel (NWK) estimator is one of the most important nonparametric kernel estimators that is often used in regression models with a fixed bandwidth. In this article, we consider the four new Proposed Adaptive NWK Regression Estimators (Interquartile Range [IQR], Standard Deviation [SD], Mean Absolute Devotion, and Median Absolute Deviation) rather than (Fixed Bandwidth, Adaptive Geometric, Adaptive Mean, Adaptive Range, and Adaptive Median). The outcomes in both simulation and actual data in leukemia cancer showed that the four new Aadaptive NWK Estimators (IQR, SD, Mean Absolute devotion, and Median Absolute Deviation) are more effective than the kernel estimations with fixed bandwidth in previous studies based Mean Square Error Criterion.

Keywords: Nonparametric regression, kernel regression, new aadaptive nadaraya-watson estimators, leukemia cancer, acute myeloid leukemia

INTRODUCTION

Nonparametric regression models aim to precisely determine the relationship between explanatory and response variables in various statistical situations; however, they are substantially less successful than parametric approaches when fitting a normal distribution. On the opposite side, nonparametric models are highly efficient in populations that do not fit normal distribution.[10] They may be used to a wide range of data types such as ordinal, nominal, ratio, and interval data.[17] Suppose the regression model for a given data points $(X_i, Y_i)_{(i=1)}^{n} \in R$ is

$$Y_i = f(X_i) + \varepsilon_i; i = 1,2,3,\ldots, \quad (1`)$$

Where ε_i is an observation error with zero mean variance σ^2 and m is unknown regression function.

Smoothing is a significant aspect in the nonparametric regression process and is a technique for expressing the dependent variable's pattern.[11] There have been four main factors for doing a nonparametric regression approaches for fitting data depending on.[17] First, to know the relationship between predictor and response variable. Second, to enable predictions of future findings without using a fixed parametric model. Third, by studying the results of individual locations, it offers a method for detecting spurious findings. Finally, working to solve the absent values by replacing or merging neighboring values of the independent variables X.[18]

Kernel Regression

Assume that x_1, x_2,..., xn from a random variable describe a sample of size (n) with density f(x).[1] bring the kernel density function of f(x) at point x:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \quad (2)$$

Kernel regression (Nadaraya-Watson Estimator) was established by Nadaraya in 1965 and Watson in 1964 which is

Corresponding Author:
Hazhar T. A. Blbas,Department of Statistics, College of Administration and Economics, Salahaddin University-Erbil, Kurdistan Region, Iraq.
E-mail: hazharstat@gmail.com

Received: September 4, 2021
Accepted: October 1, 2021
Published: October 30, 2021

DOI: 10.24086/cuesj.v5n2y2021.pp32-37

one of the most frequently used technique in nonparametric.[5,6,12] Both Nadaraya and Watson indicated the general estimator of $\hat{f}(x)$ in nonparametric regression related to smoothing bandwidth (h) and kernel (k) in bellow formula.

$$\hat{f}(x)=\frac{\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)y_i}{\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)} \tag{3}$$

$$\hat{f}(x)=\sum_{i=1}^{n}w_i y_i \tag{4}$$

$$u=\left(\frac{x_i-x}{h}\right) \tag{5}$$

The smoothing or bandwidth parameter h is used to control the smoothness of the approximate graph, and the kernel weights as identified by

$$w_i=\frac{\sum_{i=1}^{n}\left(\frac{x_i-x}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{x_i-x}{h}\right)} \tag{6}$$

The most important factor to consider in nonparametric regression is choosing bandwidth and kernel function. On the other hand, the selection of bandwidth is much more essential than the choice of kernel function on estimation.[8] The smoothed function can be expressed by scanning each data point with a weighted kernel function and then evaluating the input at each point. The kernel function is penalized based on its range from the centered location, and the degree of this penalty is defined by the bandwidth.[8] A narrow (small) bandwidth of h results in a wiggly curve and a wealth of noise in estimation, whereas a vast (big) bandwidth of h results in a flat curve and over-smoothed curves in estimation.

One of the assumptions of kernel density function is a symmetric that is often applied with a standard normal density.[13] The kernel's functions of K can be used to one of several frequently used functions, namely Epanechnikov, Triangle, Quartic, Gaussian, Uniform, and Tricube (Triweight).[3] Gaussian is a common and practical Kernel Density Function[2] which is used in this paper as shown below.

$$K(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2};u\in[-\infty,\infty] \tag{7}$$

$$\hat{f}_h(x)=\frac{\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}exp-\frac{1}{2}\left(\frac{x_i-x}{h}\right)^2 y_i}{\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}exp-\frac{1}{2}\left(\frac{x_i-x}{h}\right)^2} \tag{8}$$

If you already have over one predictive variable, lengthy distributions, or multi modal distributions, Fixed Nadaraya Watson (FNW) is not always the best option.[9] In this situation, we can use the Variable Nadaraya-Watson Kernel (VNWK) estimator with a variable bandwidth i(x_i) as seen below:

$$\hat{f}_{VNW}(x)=\frac{\sum_{i=1}^{n}K\left(\frac{x_i-x}{h(x_i)}\right)\frac{y_i}{h(x_i)}}{\sum_{i=1}^{n}K\left(\frac{x_i-x}{h(x_i)}\right)\frac{1}{h(x_i)}} \tag{9}$$

[14] showed a formula to compute $h(x_i)$ in 1982,

$$h(x_i)=\frac{h}{\sqrt{f(x_i)}} \tag{10}$$

While $f(x_i)$ is determined by the kernel function estimator which is a probability density function of x_i.

A new algorithm developed for the Abramson design estimator by Silverman in 1986, which he called an Aadaptive NWK (ANWK) function estimator. For a fixed h, the previous Kernel function estimator was used as a first stage,[4] which is represented by $\hat{f}(x_i)$ and he established local h factor θ_i as:

$$\theta_{G_i}=\left[\frac{\hat{f}(x_i)}{G}\right]^{-\alpha} \tag{11}$$

While G represents the geometric mean of $\hat{f}(x_i)$ with G≠0 and a illustrates the sensitivity parameter between 0 and 1 ($0\leq\alpha\leq1$).[14] Selected the value of $\alpha=0.5$ because its gives an accurate predictive results. Moreover then, Silverman provided an adaptive h as seen below:

$$h(x_i)=\theta_{G_i}h \tag{12}$$

In 2010,[15] it is introduced a change to the ANW approach that they have used arithmetic mean $\bar{x}$ of $\hat{f}(x_i)$ instead of using G for figuring out the h factor in NWK estimator.

$$\theta_{\bar{x}_i}=\left[\frac{\hat{f}(x_i)}{\bar{x}}\right]^{-\alpha} \tag{13}$$

In 2014,[7] it modified the ANW approach that they have used range R of $\hat{f}(x_i)$ instead of using G or $\bar{x}$ for figuring out the h factor in NWK estimator.

$$\theta_{R_i}=\left[\frac{\hat{f}(x_i)}{R}\right]^{-\alpha} \tag{14}$$

In 2019,[16] it proposed another change for the ANW approach that he used median instead of using geometric, mean, or range for calculating the h factor in NWK estimator.

$$\theta_{Me_i}=\left[\frac{\hat{f}(x_i)}{Me}\right]^{-\alpha} \tag{15}$$

New Proposed NWK function estimator

In this study, a new changes for the Adaptive Nadaraya-Watson approach was proposed, which used four different statistical techniques such as Interquartile Range (IQR), Standard Deviation (SD), Mean Absolute Deviation (MAD), and Median Absolute Deviation (MeAD) of $\hat{f}(x_i)$ instead of using geometric mean, arithmetic mean, range, or median for figuring out the h factor in NWK estimator.

$$\theta_{P_i}=\begin{cases}\left[\dfrac{\hat{f}(x_i)}{IQR}\right]^{-\alpha}\\ \left[\dfrac{\hat{f}(x_i)}{SD}\right]^{-\alpha}\\ \left[\dfrac{\hat{f}(x_i)}{MAD}\right]^{-\alpha}\\ \left[\dfrac{\hat{f}(x_i)}{MeAD}\right]^{-\alpha}\end{cases} \tag{16}$$

Where IQR is an abbreviation for IQR, SD represents SD, MAD is an acronym for MAD, and MeAD stands for Median Absolute Deviation.

Mean Square Error (MSE)

In this paper, MSE is used as an estimation criterion to find out the difference between (classical) traditional and newly proposed NWK estimators. As mentioned below, the best estimator will be the one with lowest MSE value.

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{17}$$

MATERIALS AND METHODS

Leukemia cancer data are used to undertake the performance of all proposed methods such as ANW IRQ, ANW SD, ANW MAD, and ANW MeAD in real application which collected from January 2015 to December 2020 at Nanakali Hospital for Blood in Erbil City of Iraq.

Furthermore, the CD45 outcome as an explanatory variable and Platelet (PLT) as a response variable in AML type of Leukemia cancer from 30 patients in Table 1 is used to compare between proposed methods and classical methods.

Since simple regression cannot be met due to assumptions such as linearity and autocorrelation, nonparametric regression is used for both classical and proposed approaches based on MSE. Table 2 compares the findings of classical methods to a new proposed modification of the ANW kernel estimator for IQR, SD, MAD, and MeAD based on improving the prediction accuracy of the ANW kernel estimator. The proposed MeAD approach has the smallest MSE in various bandwidths and sample sizes, followed by MAD, SD, and IQR, respectively. We achieved the same results as in the simulation analysis, namely that all new proposed methods have smaller MSE than all classical methods.

Simulation Study

A simulation analysis was carried out to compare the efficiency between traditional Nadaraya-Watson and new proposed methods of ANW estimators using the R (4.0.2) language software. An explanatory variable and response variable with adding noise to an exponential wave are simulated in this non-linear regression function to identify that proposed methods outperforms classical models.

$$y_i=\exp(x_i)-x_i^3+\varepsilon_i\sim N(0,1) \tag{18}$$

Where xi was chosen at random from a uniform distribution on the interval [–2, 2] and generated different samples of size (25, 50, 75 and 150) with different fixed *h* (0.5, 0.75, 1, and 1.5). In this paper, comparison between classical methods such as FNW, VNW Geometric, ANW Mean, ANW Median, ANW Range with new proposed methods such as ANW IRQ, ANW SD, ANW MAD, and ANW MeAD kernel estimations were computed by Gaussian kernel function with 20000000 repetitions in each estimation.

Figure 1 represents the results between classical and new proposed methods at h (5, 15, and 30) for real data. On the

Table 1: AML Type of leukemia cancer from 2015 to 2020

CD45	PLT	CD45	PLT	CD45	PLT	CD45	PLT	CD45	PLT
70	17	80	54	62	13	80	89	74	61
70	51	88	31	60	15	90	6	77	16
80	7	58	16	32	62	88	137	80	155
94	79	54	40	42	71	60	2	55	24
40	44	70	95	54	19	35	58.9	70	31
85	68	74	46	86	5	43	33	80	229

Table 2: MSE values between classical methods and proposed methods of NWK estimators in real data (n=30)

h	Fixed NW	Variable NW G	ANW M	ANW Me	ANW R	ANW IQR	ANW SD	ANW MAD	ANW MeAD
5	1890.7	1751.1	1768.2	1781	1787.9	1609.8	1609.7	1574.3	1572.4
10	2071	2010.9	2017.1	2015.6	1964.9	1909.1	1827.6	1803.2	1802.8
15	2214.5	2176.2	2179.1	2179.2	2046.3	1983.4	1906	1889.1	1889
20	2297.3	2277.5	2278.5	2279.3	2111.5	2021.5	1945.7	1928.5	1927.2
25	2340.4	2331.2	2331.6	2332.1	2152.3	2029.6	1966.8	1947.9	1945.8
30	2365.4	2360.9	2361.1	2361.4	2174.6	2033.7	1978	1958	1955.5

other hand, Figures 2 and 3 illustrate the simulation results for both classical and proposed methods in nonparametric regression functions with sample size of 75 and at h (0.5, 07.75, 1, and 1.5), respectively.

Table 3 compares simulation results between all classical (traditional) methods and new proposed update of Adaptive NW kernel such as IQR, SD, MAD, and MeAD aimed at improving the prediction accuracy of ANWK estimator. According to MSE criteria, the new proposed methods outperform classical methods of using various sample sizes and initial bandwidth values.

While the new proposed IQR is better than other classical approaches, it is less effective than other proposed methods such as SD, MAD, and MeAD, respectively. On the other hand,

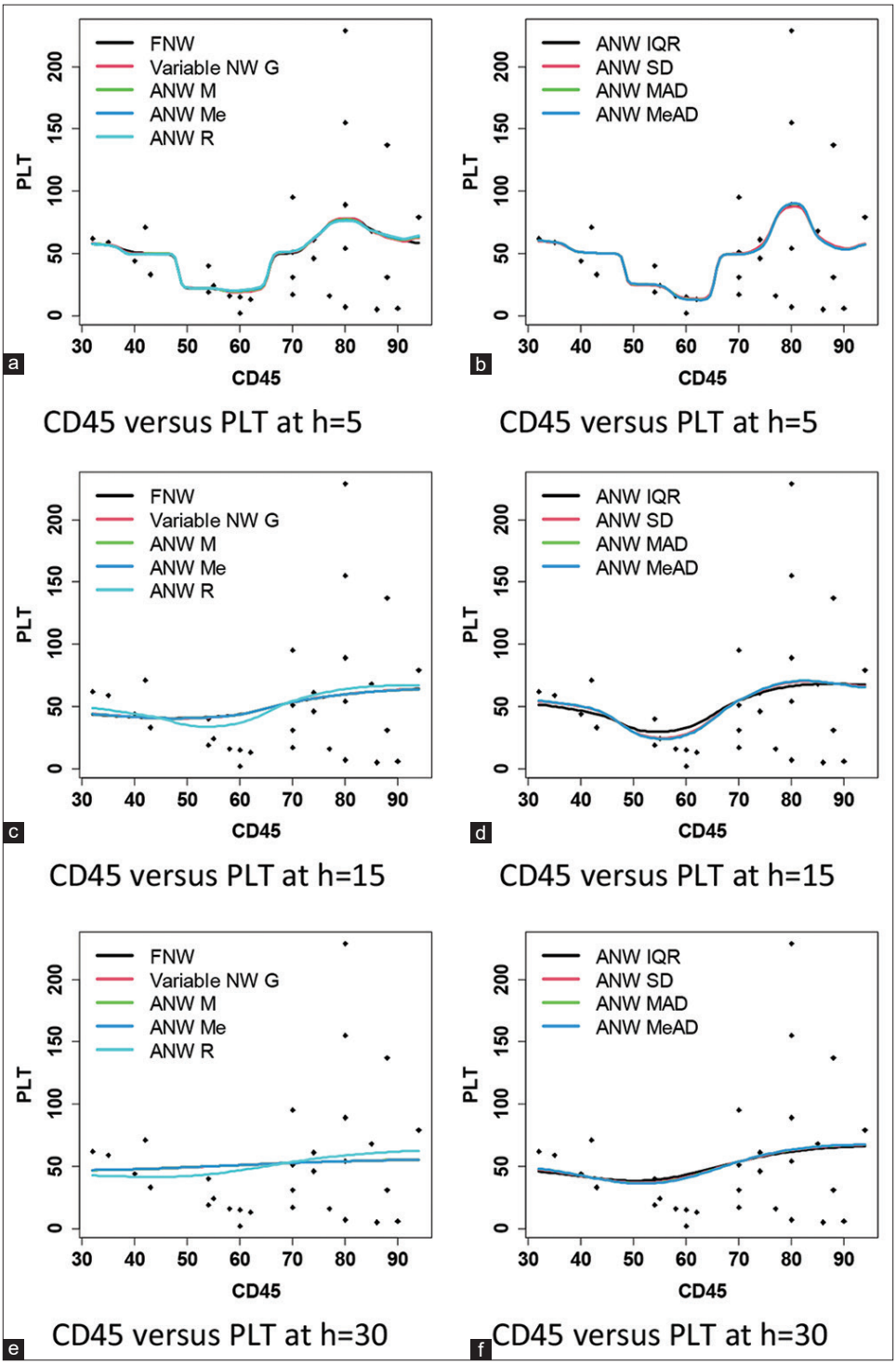

Figure 1: Left hand classical methods and right hand proposed methods at h=5, 15, and 30 for real data. (a) CD45 versus PLT at h=5, (b) CD45 versus PLT at h=5, (c) CD45 versus PLT at h=15, (d) CD45 versus PLT at h=15 (e) CD45 versus PLT at h=30 f- CD45 versus PLT at h=30

Table 3: MSE values between classical methods and new proposed methods of NWK estimators in simulation data

h	n	Fixed NW	Variable NW G	ANW M	ANW Me	ANW R	ANW IQR	ANW SD	ANW MAD	ANW MeAD
0.5	25	1.1997	1.2769	1.3439	1.3116	1.7164	1.0449	1.0124	0.9336	0.9317
	50	1.5189	1.5314	1.6769	1.4909	2.2494	1.5075	1.5106	1.4193	1.3552
	75	1.3763	1.5314	1.6357	1.4530	2.1951	1.3629	1.4166	1.3108	1.2386
	150	1.3899	1.5314	1.6141	1.3753	2.3476	1.1274	1.3872	1.2582	1.1485
0.75	25	1.7391	1.5314	1.8532	1.8887	2.1838	1.2308	1.5290	1.3999	1.3796
	50	1.9968	1.5314	2.1020	1.9050	2.5395	1.9694	1.7755	1.6786	1.6309
	75	1.9562	1.5314	2.1081	1.9022	2.5044	1.8816	1.6998	1.5897	1.5259
	150	1.9938	1.5314	2.1887	1.8970	2.7280	1.6904	1.6997	1.5504	1.4423
1	25	2.1124	1.5314	2.1651	2.1877	2.3063	1.6398	1.6593	1.5147	1.5114
	50	2.4127	1.5314	2.4600	2.3296	2.7635	2.2906	1.9823	1.8860	1.8617
	75	2.4083	1.5314	2.4799	2.3478	2.7244	2.2515	1.9349	1.8337	1.7994
	150	2.5004	1.5314	2.6277	2.4347	2.9452	2.1524	1.9254	1.7770	1.7158
1.5	25	2.4158	1.5314	2.4393	2.4446	2.4028	2.0489	1.8624	1.7373	1.7350
	50	3.0111	1.5314	3.0178	2.9693	3.0401	2.5985	2.2131	2.1163	2.1072
	75	2.9941	1.5314	3.0051	2.9563	2.9868	2.5805	2.1996	2.1043	2.0916
	150	3.1563	1.5314	3.2067	3.1444	3.1738	2.5569	2.1742	2.0404	2.0173

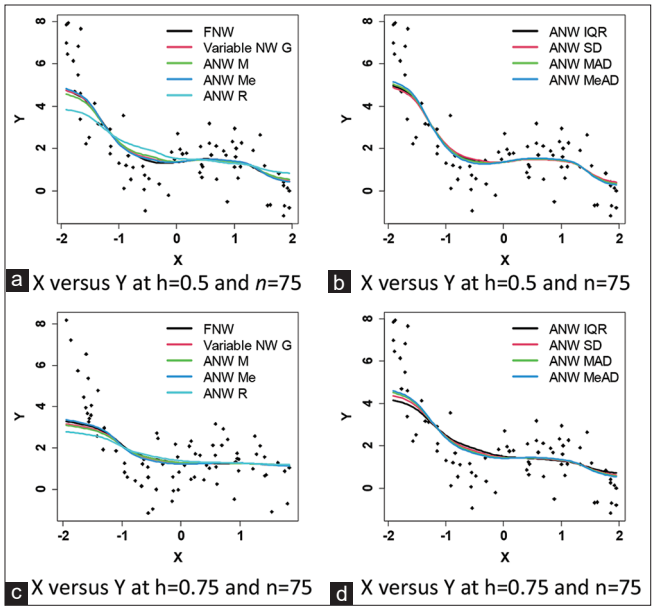

Figure 2: Left hand classical methods and right hand new proposed methods at h=0.5 and 0.75 with ample size 75, (a)-X versus Y at h=0.5 and n = 7, (b)-X versus Y at h=0.5 and n = 75, (c)-X versus Y at h=0.75 and n = 75, (d)-X versus Y at h=0.75 and n = 75

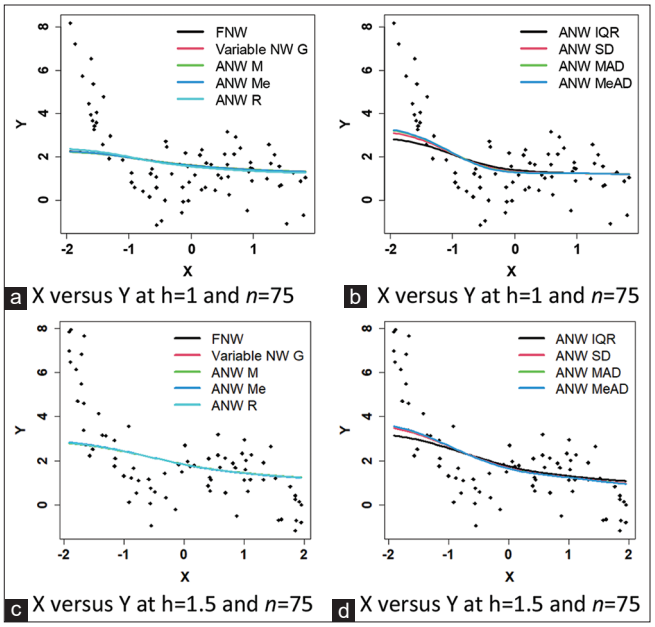

Figure 3: Left hand classical methods and right hand new proposed methods at h=1 and 1.5 with sample size 75, (a)-X versus Y at h=1 and n = 75 (b)-X versus Y at h=1 and n = 75, (c)-X versus Y at h=1.5 and n = 75 (d)-X versus Y at h=1.5 and n = 75

the proposed method using MeAD has less MSE than MAD, in turn MAD is less than SD, and SD is less than IQR in different bandwidth and sample size.

CONCLUSSION

1- The new proposed method estimator for Median Absolute Deviation (MeAD) was more reliable than any of the other classical methods for both simulation and actual data depending on MSE criteria because it is able to reduce the effect of outliers on the fitting Kernel model.

2- New proposed method estimator for MAD was more reliable than any of the other classical methods for both simulation and specific data because the absolute mean difference in MAD gives lower value of MSE.

3- New proposed method estimator for SD was more reliable than any of the other classical methods for both simulation and current data using MSE criteria.
4- New proposed method estimator for IQR was more reliable than any of the other classical methods for both simulation and real data based on MSE criteria.
5- Median Absolute Deviation (MeAD) has less MSE than MAD, in turn MAD is less than SD, and SD is less than IQR in different bandwidth and sample size in either simulation or real data.
6- In both simulation and real data, both proposed and classical methods are improved by reducing the initial bandwidth values and sample size.

REERENCES

1. M. Hollander, D. A. Wolfe and E. Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2014.
2. W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, England, New York, 1997.
3. M. Memmedli and M. Yildiz. Comparison study on smoothing parameter and sample size in nonparametric fuzzy local polynomial regression models. In: *2012 IV International Conference "Problems of Cybernetics and Informatics"* (*PCI*), 2012.
4. W. Hardle. Applied nonparametric regression. *Biometrics*, vol. 50, no. 2, p. 592, 1994.
5. A. Christmann and I. Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, vol. 13, no. 3, pp. 799-819, 2007.
6. E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, vol. 9, no. 1, pp. 141-142, 1964.
7. G. S. Watson. Smooth Regression Analysis, *Sankhya The Indian Journal of Statistics* Series *A*, vol. 26, no. 4, pp. 359-372, 1964.
8. M. P. Wand and M. C. Jones, *Kernel Smoothing*. Springer, Boston, MA, 1995.
9. H. Takeda, S. Farsiu and P. Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349-366, 2007.
10. W. L. Martínez and A. R. Martínez. *Computational Statistics Handbook with MATLAB*. Chapman & Hall, London, 2008.
11. D. Conn and G. Li. An oracle property of the Nadaraya-Watson Kernel estimator for high–dimensional nonparametric regression. *Scandinavian Journal of Statistics*, vol. 46, no. 3, pp. 735-764, 2018.
12. B. W. Silverman. *Density Estimation for Statistics and Data Analysis Estimation Density*. Kluwer Academic Publishers, London, 1986.
13. M. Hanif, S. Shahzadi, U. Shahzad and N. Koyuncu. On the adaptive Nadaraya-Watson Kernel estimator for the discontinuity in the presence of jump size. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 22, no. 2, p. 511, 2018.
14. I. S. Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, vol. 10, no. 4, pp. 1217-1223, 1982.
15. D. Li and R. Li. Local composite Quantile regression smoothing for Harris recurrent Markov processes. *Journal of Econometrics*, vol. 194, no. 1, pp. 44-56, 2016.
16. S. Demir and Ö. Toktamiş. On the adaptive Nadaraya-Watson Kernel regression estimators. *Hacettepe Journal of Mathematics and Statistics*, vol. 39, no. 3, pp. 429-437, 2010.
17. H. A. Khulood and I. Al Turk Lutfiah. Modification of the adaptive Nadaraya-Watson Kernel regression estimator. *Scientific Research and Essays*, vol. 9, no. 22, pp. 966-971, 2014.
18. T. H. Ali. Modification of the adaptive Nadaraya-Watson Kernel method for nonparametric regression (simulation study). In: *Communications in Statistics-Simulation and Computation*. Taylor & Francis Group, United Kingdom, pp. 1-13, 2019.