RESEARCH ARTICLE | DECEMBER 22 2023

Using simulations to estimate adaptive parameters in a nonparametric regression

Noor Salah Hassan; Sahera Hussein Zain Al-Thalabi

AIP Conf. Proc. 2977, 040029 (2023)
https://doi.org/10.1063/5.0182267

Using Simulations To Estimate Adaptive Parameters In A Nonparametric Regression

Noor Salah Hassan, [a] and Sahera Hussein Zain Al-Thalabi, [b]

Department of Statistics, College of Administration and Economics, Basra University, Basra, Iraq

[a] Corresponding author: noorsalahh1997@gmail.com
[b] sahera.zain@uobasrah.edu.iq

Abstract. The researcher faces several when estimating the nonparametric regression functions because the estimation methods depend on the data, as these estimates may be inaccurate, or they may not be suitable for the nonparametric model, Therefore, the aim of the research is to find the adaptive estimators in the nonparametric regression through the adaptive bandwidth method, which is known as "Goldenshluger-lepski" to increase the estimation efficiency in the nonparametric regression.. In this paper, adaptive estimations were processed in the nonparametric regression method through the use of kernel smoothing and spline. The adaptive "Goldenshluger-Lepski" was included, and to compare the estimation methods three criteria were used, namely (MSE , MAS, RMSE) to choose the best method after applying the procedure to the simulation in the R Package.

Keywords: Adaptive estimation, nonparametric regression, kernel and spline smoothing, Goldenshluger-Lepski bandwidth.

INTRODUCTION

There are many problems that the researcher faces when estimating the nonparametric regression functions because the estimation methods depend on the data, as these estimates may be inaccurate, or they may not be suitable for the nonparametric model, so the aim of this study is to find the adaptive capabilities in the nonparametric regression using modern methods. To increase the efficiency of estimation through the use of adaptive estimators in nonparametric regression smoothing . We will discuss some studies that used the adaptive method and its use in nonparametric regression, including:

The study (Hill and others, 1988) used two nonparametric adaptive procedures to apply multiple comparisons and a test of alternatives required in a one-way ANOVA model, in comparison with the parametric normal theoretical procedure, and the rank-based non-parametric procedure where these procedures are applied to lung cancer data. The results showed the superiority of the adaptive procedures Nonparametric. [6]

Astudy (2021, page and Grunewalder) presented an Adaptive estimation using the modern Goldenshluger-Lepski method to choose parameters for the statistical estimator using only the available data without making strong assumptions about the estimation. Nucleus . This method was used to address two regression problems, the kernel regression was fixed in one of them and in the other an adaptation was used. [12]

The study (Breunig and Chen, 2022) aimed to find an adaptive estimation of the minimum quadratic function in the model of non-parametric automatic variables (NPIV), which is an important problem in the optimal estimation of non-linear functions, this problem is solved through a choice based on data from Lepski type For the smoothing parameter, the results showed that the adaptive estimator of the quadratic function achieves the minimum optimum rate. [3]

Adaptive estimator in nonparametric regression: [10][1][3][4][8][11] An adaptive estimator is defined as an effective estimator for only a partially specified model (“effective” meaning that it is asymptotically equivalent to a

4th International Scientific Conference of Alkafeel University (ISCKU 2022)
AIP Conf. Proc. 2977, 040029-1–040029-9; https://doi.org/10.1063/5.0182267
Published by AIP Publishing. 978-0-7354-4776-9/$30.00

non-parametric "likelihood Maximum" local probability estimator Applicable), or a model whose distribution is unknown, so adaptive estimation aims to build estimations entirely based on data without making strong assumptions about the estimation. Nonparametric regression is also a form of regression analysis and a common and flexible tool for data analysis and modeling of the non-linear relationship between dependent and explanatory variables. , that is, it depends mainly on the data, Where the objective of the nonparametric regression is to estimate the regression function without dependence or having prior knowledge of its functional form , and using adaptive methods, Classical methods can also be modified to be as robust as non-parametric methods . Studies to build a method for selecting data-based smoothing parameters in order to obtain adaptive estimates. The first adaptive estimate was proposed by (lepski 1990) and was developed in (1992) and its goal was to build capabilities from the data in the best possible way and reduce the risk of estimation, the adaptive methods in regression The non-parametric is strong in efficiency as it cannot be outweighed by any non-adaptive method, as the exact adaptive procedure will work well with the data. So the adaptive approach is mainly divided into two types, The adaptive procedure for estimating unknown parameters is such as in a nonparametric regression, or the use of Data to determine the appropriate statistical procedure, the adaptive non-parametric approach on the one hand is estimating the parameters from the sample, or data-driven methods may be the best and most, The first to suggest this approach (Randles and Hogg). So the main purpose of adaptive approaches may be to provide a relatively easy alternative to parameterization without much effort on how to choose one from a variety of methods, and to facilitate the decision on the use of the appropriate technique. Adaptive approaches can perform better based on the available information in terms of achieving the desired combination of robustness and efficiency. over the past ten years. Adaptive-order tests show that adaptive actions Adaptive method can increase the power of tests, If the distribution of random error is abnormal, the power of classical tests is much lower than adaptive tests.

The formula for nonparametric regression is as follows:

$y_i = m(xi) + \varepsilon i$ .... (1) , i=1,2,…,n , ε~N(0,σ^2)

Yi: the response variable, m (xi): the unknown function to be estimated, xi: the explanatory variable , εi: the values of the random variable, which is white noise that is normally distributed.

The adaptive estimator for the parameter vector is as follows [15] :

$$\hat{\theta}(x) = \hat{\theta}_k(x) = \left(\theta_k^1(x), \dots\dots, \theta_k^p(x)\right)^T \quad \dots.(2) \qquad k = 1, \dots., p$$

$\theta_k^1, \dots\dots, \theta_k^p$:Unknown parameter, θ is estimated based on sample observations $(xi\ , yi)$

KERNEL SMOOTHERS

The positional polynomial regression smoother (LLS) is one of the best smoothing methods because it deals with static and random models, and it is sometimes called the weight or window function, as this function is continuous and symmetric, its integral is equal to the integer one, when (the bandwidth) is small very . [10]

The formula for smoothers is as follows

$$\widehat{m_h}(x) = \frac{\sum_{i=1}^{n} yi\ k\ (x - Xi)/h}{\sum_{i=1}^{n} k\ (x - Xi)/h} \quad \dots.(3)$$

$$w_i(x) = \frac{\frac{k(x - Xi)}{h}}{\sum_{i=1}^{n} k\ (x - Xi)/h} \quad \dots..(4)$$

$\frac{\sum_{i=1}^{n} k(x-xi)}{h}$: represents the endodontic function, $w_i(x)$: represents the weight function and one of its conditions is positive, h: represents the smoothing parameter (the bandwidth) in the estimator ($m(x)$). If its value is large, the function is smooth, and if its value is small, the function is not smooth. [1]

The Gasser-Müller (GM) smoother is one of the most widely used gradient smoothing tools. The Gasser-Müller estimator which is a modification of the Priestley-chao, estimator is used to construct nonparametric estimates of the regression function,), a new type of kernel. [4]

Its general form is as follows. [4]

$$\widehat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{h}\right) du \quad \dots \quad (5) \quad s_0 = 0 \quad , \quad s_n = 1$$

$$s_i = \frac{x_i + x_{i+1}}{2} \qquad , x_i \le s_i \le x_{i+1}$$

Also, the nearest neighbor smoother (K-NN) depends on calculating the Euclidean distance between each point and the point closest to it. If the data are close to each other, the distance will be small and vice versa. [9]

So, its general form is as follows:

$$\hat{m}_k(x) = \frac{\frac{k(xi-x)}{k_l}}{\sum_{i=1}^{n}\frac{k(xi-x)}{k_l}} \qquad k_l \to \infty \ldots (6)$$

$$, k_l = d(i,j) = \sqrt[2]{\sum_{i=1}^{n}(xi-xj)^2} \ldots \ (7)$$

k_l:represents the Euclidean distance between x, k and : xi ,xj data points

SPLINE SMOOTHER

depend on the sum of the squares of the error as used when the regression line is divided into pieces, as the explanatory variable x with period (a,b) is divided and the lines cut are called slide nodes so that smoothing the slides overcomes the problem of choosing a node and from During the identification of new nodes or changing the existing nodes, they are divided into linear spline (SPL) and cubic spline (SPC). [2][9]

$$S(m) = \sum_{i=1}^{n}\left(yi - \hat{m}(xi)\right)^2 + \lambda\int_a^b [\hat{m}''(x)]^2\,dx \ \ldots \ (8) \ , \quad \lambda > 0$$

Whereas

$\sum_{i=1}^{n}\left(yi - \hat{m}(xi)\right)^2$: It represents the sum of the squares of the error, $\hat{m}''(x)$: Represents the second derivative of the bootstrap function , λ : Represents the penalty factor indicating the width of the appropriateness quality package represented by $\sum_{i=1}^{n}\left(yi - \hat{m}(xi)\right)^2$ And the smoothing of appreciation represented by $\int_a^b [\hat{m}''(x)]^2\,dx$

Goldenshluger-Lepski adaptive bandwidth extends Lepski's method for performing adaptation across multiple parameters .This method has been used in different contexts as it was used for the first time in a multidimensional white noise model. As it has been widely used in recent studies of non-parametric estimation, the idea of this method for adaptive non-parametric estimation is to choose an estimator that reduces the sum of the unknown bias factor of variance. [8][12]

The Goldenshluger-lepski formula is as follows [5] :

$$\hat{h}(x_i) = \arg min_{h\in Ḩ_n}\left\{\hat{A}(h,x_i) + \hat{V}(h,x_i)\right\} \ \ldots (9)$$

$$\hat{A}(h,x_i) = max_{h'\in Ḩ_n}\left([\hat{m}_{h'}(x_i) - \hat{m}_{h\vee h'}(x_i)]^2 - V(h',x_i)\right) .. (10)$$

$$\hat{V}(h,x_i) = k\sigma^2\frac{\ln n}{n\hat{\varphi}(h)} \ , h \neq 0 .. (11)$$

K : represent a constant that does not depend on h , $\hat{m}_h(x_i)$: function estimator , $Ḩ_n$: Represents a set of smoothing parameter (bandwidth) .

$\hat{V}(h,x_i)$: Represents an empirical analogue of variance , $\hat{A}(h,x_i)$: Represents an approximation of the term bias

In order to estimate the regression curve, there are several criteria that are relied upon in the differentiation, and among these criteria are the mean absolute error squares (MAS), the roots mean squares error (RMSE), and the mean squared error (MSE) standard [9][14]. The function was used Endodontic (Epanchnickov) and adaptive bandwidth (Goldenshluger-lepski) on the experimental side.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{m}(x)| \ldots \qquad (12)$$

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{m}(x)\right)^2} \quad \ldots \ (13)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}(x))^2 \ldots \ (14)$$

STATISTICAL ANALYSIS

The analysis of this study is carried out using simulation, as it is known as a method that includes the use of a theoretical mathematical model and similar to the real model that represents the studied problem. Simulation

experiments were carried out using three sample sizes (n = 30, 60, 100) and with a frequency of 500 for each experiment. The nonparametric methods will be compared, and two models were used in the simulation.

first model $m(x_i) = 1 + 0.8\mathrm{e}^{-200.(-0.5+x)^2} + 2x^2$

second model $m(x_i) = \mathrm{Sin}(4\pi x)$

The variables (independent and random error) were generated. The random errors are normally distributed with a mean of zero and variance σ^2, The nonparametric explanatory variable Xi is generated according to the standard normal distribution.

TABLE 1. The first model, (RMSE, MSE, MAE) criteria for the first model according to the different sample

			RMSE			
σ^2	n	ALLS	AGM	KNN	ASPL	ASPC
$\sigma^2 = 0.5$	30	0.516265	0.930748	0.893772	0.928897	0.932278
	60	0.558081	0.861925	0.809613	0.842828	0.868399
	100	0.51343	0.759951	0.803071	0.833716	0.818787
$\sigma^2 = 1$	30	1.304713	1.293053	1.391125	1.297777	1.300542
	60	1.033427	1.288025	1.256423	1.277987	1.282404
	100	1.134386	1.143293	1.373155	1.14562	1.155014
$\sigma^2 = 1.5$	30	1.757319	1.737183	2.094056	1.742326	1.738227
	60	1.661094	1.739168	2.06102	1.716176	1.72202
	100	1.540448	1.666366	1.550798	1.664799	1.680947
				MSE		
$\sigma^2 = 0.5$	30	0.266529	0.866292	0.798829	0.862849	0.869143
	60	0.311454	0.742915	0.655473	0.710359	0.754118
	100	0.263611	0.577525	0.610768	0.695083	0.670413
$\sigma^2 = 1$	30	1.702275	1.671985	1.93523	1.684224	1.69141
	60	1.046111	1.659008	1.067971	1.633252	1.644561
	100	1.309619	1.307119	1.885555	1.312445	1.334058
$\sigma^2 = 1.5$	30	3.088172	3.024758	4.385072	3.035701	3.021432
	60	2.759233	3.024707	4.247802	2.945262	2.965353
	100	2.69107	2.776777	2.404973	2.771554	2.825582
				MAE		
$\sigma^2 = 0.5$	30	0.412093	0.718155	0.715842	0.730157	0.727187
	60	0.471807	0.651849	0.627654	0.639025	0.652287
	100	0.39879	0.63034	0.601175	0.68473	0.673663
$\sigma^2 = 1$	30	0.997865	1.001918	1.116884	1.002198	1.014442
	60	0.955424	0.963265	1.013914	0.996628	0.993963
	100	0.851318	0.850425	0.826585	0.849166	0.867145
$\sigma^2 = 1.5$	30	1.343314	1.409301	1.678374	1.380658	1.379656
	60	1.281802	1.288404	1.525217	1.289815	1.332256

sizes and levels of variation

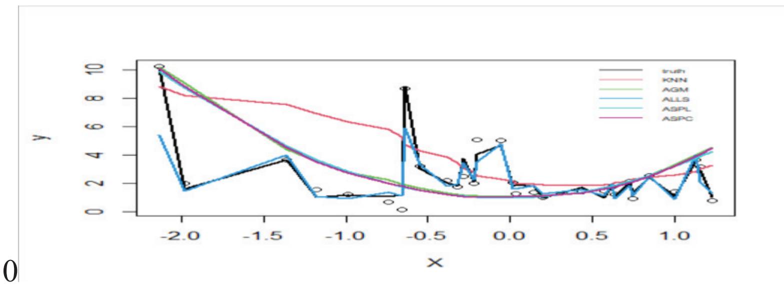
$$e_i \sim N(0, \sigma^2)$$
$$X_i \sim N(0,1)$$

Source/ From the (R.4.1.2) Package using simulation method

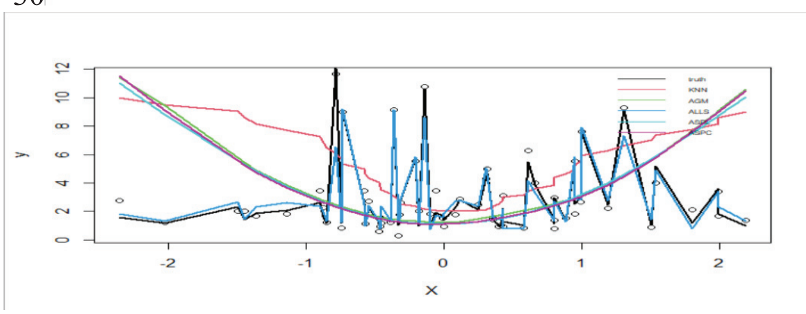
Explanation of Table 1. for the first model

1-The results showed, depending on the comparison criteria (RMSE) and (MSE) when the sample size is (n = 30, 60, 100) and with a level of variance ($\sigma^2 = 0.5$) that the best adaptive estimator is (ALLS), but when the level of variance is ($\sigma^2 = 1$,1.5) and sample size (n = 30), then the best adaptive estimator is ((AGM). As for (n=60 , $\sigma^2 = 1.5$) the best adaptive estimator is (ALLS), then the adaptive estimator (ASPL) .

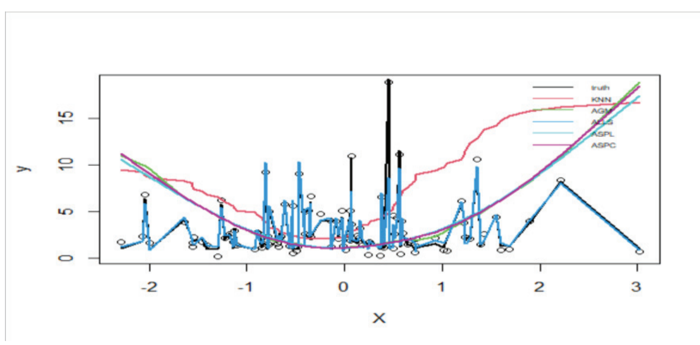
2-The results showed that, depending on the comparison standard (MAE), when the sample size is (n=30,60,100) and with the level of variance $\sigma^2 = 0.5$, the estimator is (ALLS), then the estimator is (KNN), but when the variance level is ($\sigma^2 = 1$) At the sample size (n=30,60), the best adaptive estimator is (ALLS), followed by the adaptive estimator (AGM), and when the level of variance is ($\sigma^2 = 1.5$) at the sample size (n=30), the best estimator It is an adaptive estimator (ALLS), followed by an adaptive estimator (ASPC).



n=30



n=60



n=100

FIGURE1. The adaptive nonparametric capabilities of the first model when the sample size is (30, 60, 100)

TABLE 2. For the second model, (RMSE, MSE, MAE) criteria for the second model according to the different sample sizes and levels of variance

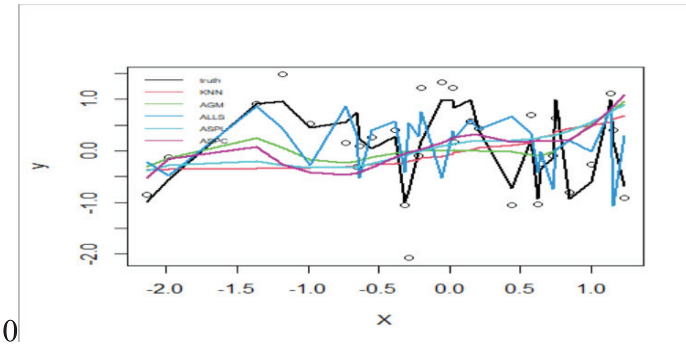
RMSE						
σ^2	n	ALLS	AGM	KNN	ASPL	ASPC
$\sigma^2 = 0.5$	30	0.903092	1.064953	0.930679	0.918608	0.907494
	60	0.813945	0.97887	0.879232	0.872201	0.878572
	100	0.626729	0.872554	0.837377	0.832611	0.833826
$\sigma^2 = 1$	30	1.24336	1.223214	1.367458	1.252031	1.250252
	60	1.220666	1.216059	1.331066	1.218084	1.222047
	100	1.156415	1.16284	1.139285	1.157729	1.163883
$\sigma^2 = 1.5$	30	1.641151	1.664737	1.97255	1.658907	1.665563
	60	1.623751	1.633353	1.82305	1.651294	1.660628
	100	1.551985	1.558501	1.680566	1.555571	1.56645
MSE						
$\sigma^2 = 0.5$	30	0.815575	1.134125	0.866163	0.84384	0.823546
	60	0.662506	0.958186	0.773049	0.760734	0.77189
	100	0.39279	0.761351	0.7012	0.693241	0.695265
$\sigma^2 = 1$	30	1.545945	1.496253	1.869941	1.567581	1.563131
	60	1.490026	1.478799	1.771737	1.483729	1.493399
	100	1.337296	1.352198	1.297971	1.340336	1.354624
$\sigma^2 = 1.5$	30	2.751423	2.771349	3.890955	2.751971	2.7741
	60	2.636569	2.667842	3.32351	2.726772	2.757686
	100	2.408658	2.428925	2.824301	2.419801	2.453765
MAE						
$\sigma^2 = 0.5$	30	0.682166	0.811518	0.773593	0.745982	0.729909
	60	0.620493	0.776988	0.719946	0.71007	0.714817
	100	0.506783	0.703813	0.684112	0.676382	0.677996
$\sigma^2 = 1$	30	0.957045	0.975529	1.029926	0.964151	0.948
	60	0.940787	0.953088	0.996644	0.945239	0.946551
	100	0.918078	0.91841	0.929677	0.91755	0.924274
$\sigma^2 = 1.5$	30	1.275417	1.318811	1.491913	1.271999	1.272058
	60	1.230053	1.228722	1.411986	1.236947	1.225396
	100	1.209956	1.196204	1.365384	1.215266	1.22144

Source/ From the (R.4.1.2) Package using simulation method

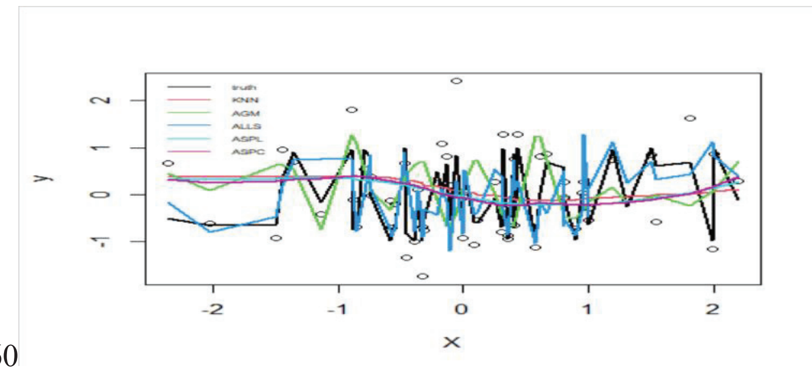
Explanation of TABLE 2. for the second model

1-The results showed, based on the comparison criteria (RMSE) and (MSE) when the sample size is (n = 30, 60, 100) and with a level of variance ($\sigma^2 = 0.5$) that the best adaptive estimator is (ALLS), followed by the second rank estimator (ASPL), But when the level of variance is ($\sigma^2 = 1$)at the sample size (n = 30, 60), the best adaptive estimator is (AGM), and when the level of variance is ($\sigma^2 = 1.5$)At the sample size (n=30, 60, 100), the best adaptive estimator is (ALLS), followed by (ASPL) estimator.

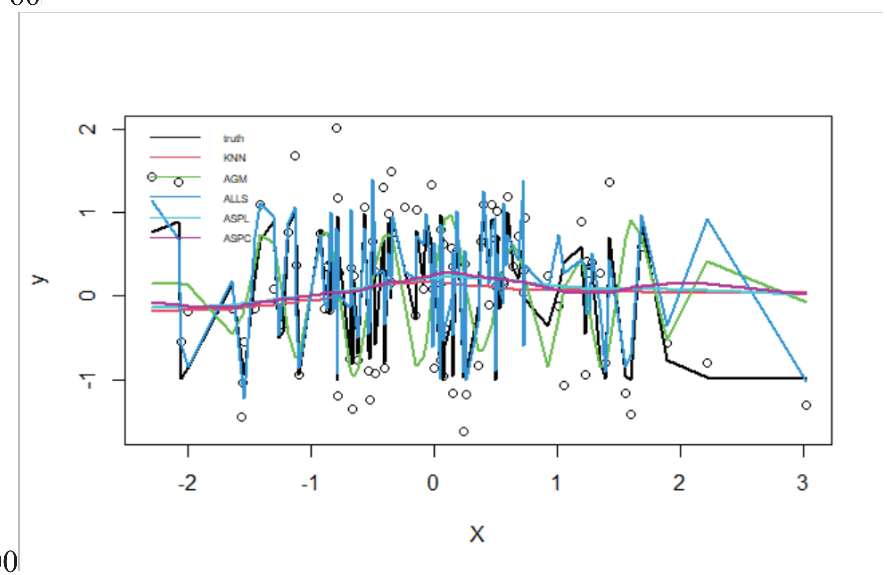
2-The results showed that, depending on the comparison standard (MAE), when the sample size is (n=30,60,100) and with a level of variance $(\sigma^2 = 0.5)$the best adaptive estimator is (ALLS), followed by an estimator (ASPL), but when the variance level is $(\sigma^2 = 1)$At the sample size (n=30), the best estimator is (ASPC), followed by the adaptive estimator (ALLS). The level of variance is $(\sigma^2 = 1.5)$ when the size is (n=30), then the best estimator is (ASPL), followed by an estimator (ASPC) and the size is (n=60), the best estimator is (AGM), then the next estimator (ASPC). At a sample size (n=100), the best estimator is (AGM), followed by an estimator (ALLS).



n=30



n=60



n=100

FIGURES 2. The adaptive nonparametric capabilities of the second model when the sample size is (30, 60, 100)

CONCLUSIONS

1. When implementing simulation experiments using three sample sizes (n = 30, 60, 100) and with a frequency of 500 for each experiment and depending on the comparison criteria at a level of variance (σ^2= 0.5), it was found that the best estimator of the first nonparametric model is that the estimator (ALLS) is the best estimator, then It is followed by the estimator (KNN).
2. But when the level of variance is (σ^2=1.5) at the sample size (n=30), the best estimator for the first model is (AGM), followed by (ASPC) estimator, as the values of the criteria (RMSE), (MSE) and (MAE) are less. With increasing sample sizes and for all estimators used, and increasing the values of (RMSE), (MSE) and (MAE) for all estimators with increasing values of residual variance.
3. As for the second model, the results showed, depending on the comparison criteria (RMSE) and (MSE), when the sample size (n = 30,60) at a level of variation (σ^2= 0.5, 1) that the estimator (ALLS) is the best estimator, then follows 2nd place estimator (ASPL).
4. But when the level of variance is (σ^2=1) at the sample size (n=30), the best estimator is (ASPC), followed by the estimator (ALLS), and in general the best adaptive estimator for the second nonparametric model is (ALLS), followed by (ASPL) estimator adaptive ,
5. Finally, we can say that the best estimation adaptive method for the three criteria and by increasing the sample sizes at three different levels of variance was the ALLS method, which represents the smoothing of the adaptive local polynomial regression.

ACKNOWLEDGMENTS

The authors would like to thank Basra University (www.uo Basra.edu.iq) Basra - Iraq for its support in the present work.

REFERENCES

1. -Ali, Noor Abdul-Karim (2021) “Comparing some of the traditional and immunized nonparametriccapabilities of the nonparametric regression model with application”, Master’s thesis, College of Administration and Economics, University of Basra
2. Aydin , D. & Memmedli2 , M. & Omay2, R. (2013) " Smoothing Parameter Selection for Nonparametric Regression Using Smoothing Spline" EUROPEAN JOURNAL OF PURE AND APPLIED MATHEMATICS Vol. 6, No. 2, 222-238 ISSN 1307-5543
3. -Breunig, C. & Chen, X. (2022) " simple adaptive estimation of quadratic functionals in nonparametric iv models" mathematics statistics theory , version, v2
4. -Fan, J. & Gijbels, L. (2017) " Local Polynomial Modelling and Its Applications ", MONOGRAPHS ON STATISTICS AND APPLIED PROHABILITY , First edition 1996, Taylor & Francis Group journal
5. -Chagnya, G. & Roche, A. (2016) " Adaptive estimation in the functional nonparametric regression model " Journal of Multivariate Analysis ,Volume 146, April
6. Hill ,N. J. , Padmanabhan, A. R. , Puri, M. L. (1988) " Adaptive Nonparametric Procedures and Applications " Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 37, No. 2,1988
7. Lacour, C.& Massart, P. (2016)" Minimal penalty for Goldenshluger–Lepski method" Stochastic Processes and their ApplicationsVolume 126, Issue 12, December
8. Lederer, J.& YU , L.& GAYNANOVA , I. (2019) Oracle Inequalities for High-dimensional Prediction, Vol.25 • No. 2 Bernoulli Society for Mathematical Statistics and Probability.
9. Matta, Nour Sabah and Al-Safawi, Safa Younes (2011) “Estimation of nonparametric regression functions using some improvement”, Iraqi Journal of Statistical Sciences, Vol. 2011, No. 20
10. Muhammad, Abdul-Hussain Muhammad (2011) “Using the estimator ((Nadaraya-Watson)) kernel in estimating the nonparametric regression function, Al-Qadisiyah Journal for Administrative and Economic Sciences, Volume 13, Issue 1
11. -O'Gorman, T. W. (2004) Applied Adaptive Statistical Methods Tests of Significance and Confidence Intervals , by the American Statistical Association and the Society for Industrial and Applied Mathematics.

12. page, S.& grunewalder , S. (2021) The Goldenshluger–Lepski Method for Constrained Least-Squares Estimators over RKHSs , Journals Project Euclid , Bernoulli 27(4): 2241-2266 , November Volume 27 , Issue 4
13. Serdyukova , N. (2012) Adaptive estimation in regression and complexity of approximation of random fields , Statistics Theory (math.ST); Probability (math.PR)
14. Tali, Nada Hussein and Taher, Ahmed Shaker Mohammed Taher (2022) "Using kernel regression in estimating the coefficients of the regression model with random error limits that are self-correlated with practical application", Journal of Administration and Economics - Al-Mustansiriya University, issue 132, pages 234-247
15. Zhang , Z. (2014) Adaptive Robust Regression Approaches in data analysis and their Applications , PhD thesis , University of Cincinnati ,ProQuest Publishing,