Smoothing parameter selection in kernel nonparametric regression using bat optimization algorithm

Marwah Yahya Mustafa¹ and Zakariya Yahya Algamal²
¹Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
²Department of Statistics and Informatics, University of Mosul, Mosul, Iraq
¹E-mail: marwa.yehea1987@gmail.com
²zakariya.algamal@uomosul.edu.iq

Abstract: In the context of kernel nonparametric regression, the curve estimation is fully depending on the smoothing parameter. At this point, the nature-inspired algorithms can be used as an alternative tool to find the optimal selection. In this paper, a bat optimization algorithm method is proposed to choose the smoothing parameter in Nadaraya-Watson kernel nonparametric regression. The proposed method will efficiently help to find the best smoothing parameter with a high prediction. The proposed method is compared with four famous of prediction capability.

1. Introduction

The nonparametric regression model (NRM) and methods of estimation have been developed mainly in the last years [1, 2]. “Kernel regression estimates are one of the most popular nonparametric estimates. In a univariate case, these estimates depend on a bandwidth, which is a smoothing parameter controlling smoothness of an estimated curve and a kernel which is considered as a weight function [3-5].

The choice of the smoothing parameter is a crucial problem in the kernel regression. The literature on bandwidth selection is quite extensive, such as [1, 3, 6-22].

In this paper, a bat optimization algorithm method, which is a natural-inspired continuous algorithm, is proposed to choose smoothing parameter in Nadaraya-Watson kernel nonparametric regression. The proposed method will efficiently help to find the best smoothing parameter with a high prediction. The superiority of the proposed method in different simulated examples and a real data application is proved.

This paper is organized as follows. The description of the Nadaraya-Watson kernel nonparametric regression and the smoothing parameter selection are covered in Section 2. The details of the bat optimization algorithm are covered in Section 3. Section 4 contains the details of our proposed method. The illustration of the proposed method through simulation studies and through real data application is given in Sections 5 and 6. In section 7, the conclusion is covered.

2. Smoothing parameter selection

The nonparametric regression model, often estimated by estimators of the Nadaraya–Watson type, forms an attractive framework for diverse areas such as engineering, econometrics, environmetrics, social sciences, and biometrics [5].

In NRM, we have a set of univariate observations \( \{(x_i, y_i)\}_{i=1}^{n} \in \mathbb{R} \) and the model can be defined as:

\[
y_i = m(x_i) + \varepsilon_i, \quad i = 1, 2, ..., n
\]  

(1)
where \( m(x_i) \) is the function of unknown regression and \( \varepsilon_i \) are random errors with mean equal to zero and variance \( \sigma^2 \). The nonparametric regression depend on weighted mean of the dependent variable, the weights are the distance between the observations of independent variable measured by a smoothing parameter.

One of the techniques nonparametric regression estimate is the Nadaraya-Watson (NW) kernel function estimator which is more flexible than the other nonparametric techniques and it provides an accurate predictor of observations [1, 23, 24].

The kernel estimator of \( f(x) \) at the point \( x \), in general, is defined as

\[
\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right),
\]

(2)

where \( K \) represents kernel probability density function centered at each point \( x_i \), and the smoothing parameter \( h > 0 \) is known as fixed bandwidth. The NW kernel function estimator with a fixed \( h \) is defined as the following:

\[
\hat{m}_{NW}(x_i) = \frac{\sum_{i=1}^{n} y_i K \left( \frac{x - x_i}{h} \right)}{\sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right)}.
\]

(3)

The NW kernel estimator depends on smoothing parameter \( h \) (bandwidth). It controls the amount of curve smoothing where large \( h \) value leads to a smooth density estimate [3, 19, 25, 26]. The optimal bandwidth of the NW kernel estimator is the value which minimized the mean integrated squared error and it is obtained by integration the mean squares of errors (IMSE). There are many different methods to select the value of \( h \). Among them, in 1981, Friedman and Stuetzle used this strong reparability to identify the components of the nonparametric regression model when \( h \) is unknown, and proposed a kernel-based consistent and asymptotically normal estimator [27]. In 1982, the researcher Abramson suggested the law of inverse square root to estimate \( h \) in variable kernel density function, which reduced the bias more than fixed h estimator [28]. In 1986, the researcher Silverman suggested an adaptation for the kernel function estimator by varying the \( h \) as nonparametric estimation depends on geometric mean [29]. In 1987, the researchers Scott and Terrell discussed relationship between the biased and unbiased cross-validation and the variable \( h \) used instead than the fixed \( h \) in the case of long-tail distribution [30]. There are several authors handling the problem of selection the smoothing parameter, such as [1, 3, 6-22].

3. Bat optimization algorithm

Nature has been an inspiration for the introduction of many meta-heuristic algorithms [31-34]. Swarm intelligence is an important tool for solving many complex problems in scientific research [31, 35]. Swarm intelligence algorithms have been widely studied and successfully applied to a variety of complex optimization problems. Bat algorithm (BA) proposed by Yang [36] is based on the echolocation ability of the microbats that guides them on their foraging behavior. This algorithm starts with the random initial population of bats in a n-dimensional search space where the position of the bat
i denoted by $x_i^t$ and its velocity denoted by $v_i^t$ at time $t$. Therefore, the new positions $x_i^{t+1}$ and new velocities $v_i^{t+1}$ at time step $t+1$ can be determined by

$$v_i^{t+1} = v_i^t + (x_i^t - G_{best})\delta_i,$$

$$x_i^{t+1} = x_i^t + v_i^{t+1},$$

$$\delta_i = \delta_{min} + (\delta_{max} - \delta_{min})\theta,$$

where $\theta$ is a random number in [0, 1], $G_{best}$ represents the current global optimal solution, and $\delta_i$ represents the pulse frequency emitted by bat $i$ at the current moment, where $\delta_{min}$ and $\delta_{max}$ represent the minimum and maximum values of pulse frequency, respectively. Initially, $\delta_i$ is assigned randomly for each bat which is elected uniformly from $[\delta_{min}, \delta_{max}]$ [36-41]. The corresponding position of the randomly bat is updated as

$$x_{new} = x_{old} + \epsilon L^i,$$

where $x_{old}$ represents a random solution chosen from the current best solutions, $L^i$ is the loudness and $\epsilon$ is a random vector that is drawn from $[-1,1]$. The pulse emission rate $r$ and the loudness $L$ are updated by controlling the balance between these techniques as follows

$$r_i^{t+1} = r_i^0 \times [1 - \exp(-b_1 t)],$$

$$L_i^{t+1} = b_2 L_i^t,$$

where $b_1$ and $b_2$ are constants.

4. The proposed method

The efficiency of NW kernel estimator largely depends on an appropriately choosing the smoothing parameter, $h$. As a result, it is of crucial importance selecting a suitable value of the $h$. In literature, the most widely used method for selecting $h$ is the cross-validation (CV), which is a data-driven approach [6, 30].

In this paper, a BA is proposed to determine the smoothing parameter in the NW kernel estimator. The proposed method will efficiently help to find the best value with high prediction performance. The parameter configurations for our proposed method are presented as follows.

1. The number of bats, is set to 20 and the number of iterations is $t_{max} = 150$.
2. The positions of each bat are randomly determined. The position of a bat represents the smoothing parameter, $h$. The initial positions of the bats are generated from a uniform distribution within the range [0,10].
3. The fitness function is defined as
fitness = \min \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad (10)

(4) The positions are updated using Eq. (7).

(5) Steps 3 and 4 are repeated until a \( t_{\text{max}} \) is reached.

5. Simulation results

To test how well the proposed method performs for different possible mean functions the following study design was followed. The comparisons with different used methods, CV, GCV, AIC, and plug-in method (PM) are also conducted. “Three sizes of samples are taken as: \( n = 50, 100, 150 \). In addition, the type of kernel is setting as Epanechnikov kernel type.

Case 1: In the case, we use the regression function

\[ y_i = x^3 (1 - x)^3 + \varepsilon_i \quad \mathcal{N} (0, 0.003) \] \quad (11)

The explanatory variable, \( x \) is generated from uniform distribution with the range 0 and 1.

Case 2: In the case, we use the regression function

\[ y_i = \sin(\pi x) \cos(3\pi x^3) + \varepsilon_i \quad \mathcal{N} (0, 0.05) \] \quad (12)

The explanatory variable, \( x \) is generated from uniform distribution with the range 0 and 1.

Case 3: In the case, we use the regression function

\[ y_i = x \sin(2\pi x) + \varepsilon_i \quad \mathcal{N} (0, 0.05) \] \quad (13)

The explanatory variable, \( x \) is generated from uniform distribution with the range 0 and 1.

Case 4: In the case, we use the regression function

\[ y_i = \sin(2\pi x) + \varepsilon_i \quad \mathcal{N} (0, 0.5) \] \quad (14)

The explanatory variable, \( x \) is generated from uniform distribution with the range 0 and 1.

Case 5: In the case, we use the regression function

\[ y_i = \sin(4x) + \varepsilon_i \quad \mathcal{N} (0, 0.11) \] \quad (15)

The explanatory variable, \( x \) is generated from uniform distribution with the range 0 and 1.

The generated data is repeated 500 times and the averaged integrated mean squared error (IMSE) is calculate. The results of the used methods are summarized in Tables 1 – 5.

The results presented in Tables 1 – 5 show that the BA algorithm yielded higher prediction accuracy than did CV, GCV, AIC, and PM methods for the case 1, case 2, case 3, case 4, and case 5. The proposed method, BA, has superior reduction in terms of IMSE. For example, for case 2, the reduction in IMSE using BA was 5.52%, 5.02%, 5.97%, and 7.84% compared with CV, GCV, AIC, and PM methods, respectively."
With respect to the sample size, there is a decreasing in the IMSE values when the sample size increasing. However, the performance of the proposed method, BA, is still the best among others in all cases.

Table 1: The average IMSE for the case 1

| Methods | n=50    | n=100   | n=150   |
|---------|---------|---------|---------|
| CV      | 0.3411  | 0.3225  | 0.3218  |
| GCV     | 0.3415  | 0.3582  | 0.3371  |
| AIC     | 0.5024  | 0.4735  | 0.4416  |
| PM      | 0.7274  | 0.556   | 0.5192  |
| BA      | **0.2491** | **0.2472** | **0.2157** |

Table 2: The average IMSE for the case 2

| Methods | n=50    | n=100   | n=150   |
|---------|---------|---------|---------|
| CV      | 0.3723  | 0.2362  | 0.2266  |
| GCV     | 0.3701  | 0.2518  | 0.2366  |
| AIC     | 0.4253  | 0.4044  | 0.3646  |
| PM      | 0.6808  | 0.5836  | 0.5202  |
| BA      | **0.2361** | **0.2158** | **0.2188** |

Table 3: The average IMSE for the case 3

| Methods | n=50    | n=100   | n=150   |
|---------|---------|---------|---------|
| CV      | 0.3822  | 0.3552  | 0.3535  |
| GCV     | 0.3656  | 0.3452  | 0.3435  |
| AIC     | 0.3964  | 0.3915  | 0.3543  |
| PM      | 0.5304  | 0.50067 | 0.4941  |
| BA      | **0.2831** | **0.2678** | **0.2431** |

Table 4: The average IMSE for the case 4

| Methods | n=50    | n=100   | n=150   |
|---------|---------|---------|---------|
| CV      | 0.4731  | 0.3966  | 0.4574  |
| GCV     | 0.4637  | 0.3936  | 0.4393  |
| AIC     | 0.4912  | 0.4622  | 0.4111  |
| PM      | 0.5204  | 0.5156  | 0.5301  |
| BA      | **0.3578** | **0.3361** | **0.3151** |

Table 5: The average IMSE for the case 5

| Methods | n=50    | n=100   | n=150   |
|---------|---------|---------|---------|
| CV      | 0.1721  | 0.1677  | 0.1622  |
| GCV     | 0.1608  | 0.1518  | 0.1456  |
| AIC     | 0.1674  | 0.1562  | 0.1474  |
| PM      | 0.2306  | 0.2174  | 0.2106  |
| BA      | **0.1365** | **0.1358** | **0.1277** |
6. Real application results
To investigate the performance of our proposed method in real application, 255 observations are daily taken from 2ed January/2019 to 13ed March/2020. This data is related to ASIACell Communication Company in Iraq. The response variable is the number of deals and the explanatory variable is the closed price of the share in Iraqi dinar. The estimated IMSE values and the value of the smoothing parameter, \( h \), is reported in Table 6. It is clearly seen from Table 6, that the proposed method, BA, achieved the lowest prediction error. It reduced the IMSE by 5.08\%, 4.87\%, 3.64\%, and 3.67\% than those obtained by CV, GCV, AIC, and PM, respectively. Figure 1 shows the smoothing curve of the used methods in addition to the linear model which is estimated by ordinary least square method (OLS). It is clearly seen that our proposed method give a more smooth curve than the others.

Table 6: The IMSE results of the application data

| Methods | \( h \) | IMSE   |
|---------|--------|--------|
| CV      | 1.970  | 0.8547 |
| GCV     | 1.851  | 0.7741 |
| AIC     | 1.024  | 0.6526 |
| PM      | 1.051  | 0.6628 |
| BA      | 0.611  | 0.5206 |

Figure 1: The estimated smoothing curve for the real data application over the used methods with the OLS method.
7. Conclusion
In this paper, the problem of selecting smoothing parameter in Nadaraya-Watson kernel nonparametric regression is considered. A bat optimization algorithm was proposed to choose the parameter of smoothing parameter. The results obtained from simulation and real data application demonstrated the superiority of the proposed method, BA, in terms of IMSE comparing with other competitor methods.

REFERENCES

[1] Ali T H 2019 Modification of the adaptive Nadaraya-Watson kernel method for nonparametric regression (simulation study) Communications in Statistics - Simulation and Computation 1-13
[2] Hazelton M L and Cox M P 2016 Bandwidth selection for kernel log-density estimation Computational Statistics & Data Analysis 103 56-67
[3] Schucany W R 1995 Adaptive Bandwidth Choice for Kernel Regression Journal of the American Statistical Association 90 535-40
[4] Slauoi Y 2016 Optimal bandwidth selection for semi-recursive kernel regression estimators Statistics and Its Interface 9 375-88
[5] Steland A 2012 Sequential Data-Adaptive Bandwidth Selection by Cross-Validation for Nonparametric Prediction Communications in Statistics - Simulation and Computation 41 1195-219
[6] Chen S 2015 Optimal Bandwidth Selection for Kernel Density Functionals Estimation Journal of Probability and Statistics 2015 1-21
[7] Chu C 1995 Bandwidth selection in nonparametric regression with general errors Journal of statistical planning and inference 44 265-75
[8] Chu C-K and Marron J S 1991 Comparison of two bandwidth selectors with dependent errors The Annals of Statistics 19 1906-18
[9] Dobrovidov A V and Ruds’ko I M 2010 Bandwidth selection in nonparametric estimator of density derivative by smoothed cross-validation method Automation and Remote Control 71 209-24
[10] Feng Y and Heiler S 2009 A simple bootstrap bandwidth selector for local polynomial fitting Journal of Statistical Computation and Simulation 79 1425-39
[11] Francisco-Fernández M and Vilar-Fernández J M 2005 Bandwidth selection for the local polynomial estimator under dependence: a simulation study Computational Statistics 20 539-58
[12] Gao J and Gijbels I 2012 Bandwidth Selection in Nonparametric Kernel Testing Journal of the American Statistical Association 103 1584-94
[13] Kauermann G and Opsomer J D 2004 Generalized Cross-Validation for Bandwidth Selection of Backfitting Estimates in Generalized Additive Models Journal of Computational and Graphical Statistics 13 66-89
[14] Koláček J and Horová I 2017 Bandwidth matrix selectors for kernel regression Computational Statistics 32 1027-46
[15] Lee T C and Solo V 1999 Bandwidth selection for local linear regression: a simulation study Computational Statistics 14 515-32
[16] Leungi D H Y, Marriott F H C and Wu E K H 1993 Bandwidth selection in robust smoothing Journal of Nonparametric Statistics 2 333-9
[17] Nychka D 1991 Choosing a range for the amount of smoothing in nonparametric regression Journal of the American Statistical Association 86 653-64
[18] Opsomer J D and Miller C P 2007 Selecting the amount of smoothing in nonparametric regression estimation for complex surveys Journal of Nonparametric Statistics 17 593-611
[19] Rice J 1984 Bandwidth choice for nonparametric regression The Annals of Statistics 12 1215-30
[20] Zhang Z G, Chan S C, Ho K L and Ho K C 2008 On Bandwidth Selection in Local Polynomial Regression Analysis and Its Application to Multi-resolution Analysis of Non-uniform Data Journal of Signal Processing Systems 52 263-80
[21] Zhou H and Huang X 2018 Bandwidth selection for nonparametric modal regression Communications in Statistics - Simulation and Computation 48 968-84
[22] Żychaluk K 2014 Bootstrap bandwidth selection method for local linear estimator in exponential family models Journal of Nonparametric Statistics 26 305-19
[23] Kyung Lee Y, Park B U and Su Park M 2007 A simple and effective bandwidth selector for local polynomial quasi-likelihood regression Journal of Nonparametric Statistics 19 255-67
[24] Li J and Palta M 2009 Bandwidth selection through cross-validation for semi-parametric varying-coefficient partially linear models Journal of Statistical Computation and Simulation 79 1277-86
[25] Utami T W, Haris M A, Prahutama A and Purnomo E A 2020 Optimal knot selection in spline regression using unbiased risk and generalized cross validation methods Journal of Physics: Conference Series 1446 012049
[26] Yoon Kim T and Park C 2002 Regression Smoothing Parameter Selection Using Cross Residuals Sum Communications in Statistics - Theory and Methods 31 2275-87
[27] Friedman J H and Stuetzle W 1981 Projection pursuit regression Journal of the American statistical Association 76 817-23
[28] Abramson I S 1982 On bandwidth variation in kernel estimates-a square root law The annals of Statistics 1217-23
[29] Silverman B W 1986 Density estimation for statistics and data analysis vol 26: CRC press
[30] Scott D W and Terrell G R 1987 Biased and unbiased cross-validation in density estimation Journal of the american Statistical association 82 1131-46
[31] Algamal Z Y 2018 A new method for choosing the biasing parameter in ridge estimator for generalized linear model Chemometrics and Intelligent Laboratory Systems 183 96-101
[32] Al-Thanoon N A, Qasim O S and Algamal Z Y 2018 Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification Computers in Biology and Medicine 103 262-8
[33] Al-Thanoon N A, Qasim O S and Algamal Z Y 2019 A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics Chemometrics and Intelligent Laboratory Systems 184 142-52
[34] Ismael O M, Qasim O S and Algamal Z Y 2020 Improving Harris hawks optimization algorithm for hyperparameters estimation and feature selection in v-support vector regression based on opposition-based learning Journal of Chemometrics 34
[35] Kahya M A, Altamir S A and Algamal Z Y 2020 Improving firefly algorithm-based logistic regression for feature selection Journal of Interdisciplinary Mathematics 22 1577-81

[36] Yang X-S 2010 Nature inspired cooperative strategies for optimization (NICSO 2010): Springer) pp 65-74

[37] Alomari O A, Khader A T, Al-Betar M A and Awadallah M A 2018 A novel gene selection method using modified MRMR and hybrid bat-inspired algorithm with β-hill climbing Applied Intelligence 48 4429-47

[38] Rizk-Allah R M and Hassanien A E 2017 New binary bat algorithm for solving 0–1 knapsack problem Complex & Intelligent Systems 4 31-53

[39] Zhao D and He Y 2015 Chaotic binary bat algorithm for analog test point selection Analog Integrated Circuits and Signal Processing 84 201-14

[40] Qasim O S and Algamal Z Y 2020 Feature Selection Using Different Transfer Functions for Binary Bat Algorithm International Journal of Mathematical, Engineering and Management Sciences 5 697-706

[41] Qasim O S, Al-Thanoon N A and Algamal Z Y 2020 Feature selection based on chaotic binary black hole algorithm for data classification Chemometrics and Intelligent Laboratory Systems 204