Comparison Study On Smoothing Parameter and Sample Size in Nonparametric Fuzzy Local Polynomial Regression Models

Memmedaga Memmedli[1], Munevvere Yildiz[2]
Anadolu University, Eskisehir, Turkey
[1]*mmammadov@anadolu.edu.tr*, [2]*munevvere@hotmail.com*

Abstract— **In this paper, we considered the relationship between the smoothing parameter value and sample size as a simulation study in nonparametric fuzzy local polynomial regression. For this aim, we developed fuzzy version of generalized cross-validation criteria (GCV) for selecting smoothing parameter in nonparametric fuzzy local polynomial models. Besides the local linear models, local cubic models are also used in these simulations. The appropriate smoothing parameters are selected by GCV criteria for different sample size and then performances of the models are compared using these appropriate smoothing parameters with sample sizes.**

Keywords— ***Local polynomial smoothing; fuzzy nonparametric regression; generalized cross validation; sample size***

I. Introduction

There are very few studies on fuzzy nonparametric regression models. It is necessary transform many notations and approaches of nonparametric regression models to fuzzy form for creating fuzzy nonparametric regression models. The first study of fuzzy parametric regression models is Ishibuchi and Tanaka's study [3]. Authors have suggested several fuzzy regression models by using back propagation neural network algorithm in this paper. Cheng and Lee [2] have used radial basis function neural networks in fuzzy regression analysis. The developments of nonparametric regression models have been getting more attention on fuzzy version of these models (see [4-8]). Aspects of this situation, Cheng and Lee [9] have examined fuzzy versions of k-nearest neighbor and kernel estimation methods. Wang, Zhang and Mei [10] have considered fuzzy local linear regression model and advantages according to fuzzy models of [9] of this approach have shown by using experimental way.

Cross-validation has used for bandwidth selection in these researches. But, Fan and Gijbels [4] have proposed more effective selection methods for local regression fitting. To improve fuzzy versions of these methods and use these in fuzzy local regression models are very important issues. Bandwidth size and order of local polynomial are associated with each other in local polynomial regression. So, analysis of this relation for fuzzy models is another important issue.

II. Methods

Let the univariate fuzzy nonparametric regression model is given by,

$$y = \mu(x)\{+\}\varepsilon \tag{1}$$

In this model, x is a crisp independent variable whose domain is assumed D and $\mu(x) = (c(x), \alpha(x), \beta(x))$ is a fuzzy function which is defined with *LR* fuzzy number. ε is a fuzzy error term and $\{+\}$ is an operator whose definition depends on fuzzy ranking method used.

Let $(x_i, y_i)(i = 1,2,...,n)$ be a sample of the observed crisp inputs and LR fuzzy outputs for model (1). Each y output is a *LR* fuzzy number which represented by (c_y, α_y, β_y) . The functions $c(x), \alpha(x), \beta(x)$ show center, right and left spreads of *LR* fuzzy number and they have continuous derivatives up to fourth order in the domain D . Taking $l_y = c_y - \alpha_y$ and $u_y = c_y + \alpha_y$, we can write this number like (l_y, c_y, u_y) by helping its lower, upper and center limits. Suppose that $l(x), c(x), u(x)$ functions are respectively the lower, center and upper limits of *LR* function which has derivatives up to *pth* order in the domain D. For a given neighborhood of $x_0 \in D$, these functions $l(x), c(x), u(x)$ can be approximately presented by *pth* order of Taylor polynomial [10].

$$\begin{aligned} l(x) &\approx l(x_0) + l'(x_0)(x - x_0) + ... + \frac{l^{(p)}(x_0)}{p!}(x - x_0)^p \\ c(x) &\approx c(x_0) + c'(x_0)(x - x_0) + ... + \frac{c^{(p)}(x_0)}{p!}(x - x_0)^p \quad (2) \\ u(x) &\approx u(x_0) + u'(x_0)(x - x_0) + ... + \frac{u^{(p)}(x_0)}{p!}(x - x_0)^p \end{aligned}$$

Based on Diamond`s distance [1] with using *pth* order polynomial, local estimators can be defined by locally weighted least squares method with a Gaussian kernel function *K(.)*;

$$(\hat{l}(x_0),\hat{l}'(x_0),...,\hat{l}^{(p)}(x_0))^T = H(x_0;h)l_y$$
$$(\hat{c}(x_0),\hat{c}'(x_0),...,\hat{c}^{(p)}(x_0))^T = H(x_0;h)c_y \quad (3)$$
$$(\hat{u}(x_0),\hat{u}'(x_0),...,\hat{u}^{(p)}(x_0))^T = H(x_0;h)u_y$$

where $H(x_0;h)=(X^T(x_0)W(x_0;h)X(x_0))^{-1}X^T(x_0)W(x_0;h)$ Instead of crisp number *y* in these estimators as defined in formulas used fuzzy numbers $W(x_0;h)=Diag(K_h(|x_1-x_0|),...,K_h(|x_n-x_0|))$ is an nxn diagonal matrix. Kernel function *K(.)* is selected Gaussian $K(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ function an $K_h(z)=\frac{1}{h}K\left(\frac{z}{h}\right)$ Because the local regression estimate solves a least squares problem, $\hat{\mu}(x)$ is a linear estimate. That is, for each *x* there exists a weight diagram vector $l(x)=\{l_i(x)\}_{i=1}^n$ such that $\hat{\mu}(x)=\sum_{i=1}^n l_i(x)y_i$ [5]. Here,

$$l^T(x)=e_1^T(X^TW(x;h)X(x))^{-1}X^T(x)W(x;h) \quad (4)$$
$$e_1=(1,0,0,...)$$

The *hat matrix* is the nxn matrix *L* with rows $l^T(x_i)$ which maps the data to the fitted values;

$$\hat{\mu}=(\hat{\mu}(x_1),...,\hat{\mu}(x_n))^T=Ly \quad (5)$$

and using Diamond distance [1] appropriate formulas can be obtained and the formula for GCV criteria is expressed as follow:

$$GCV\ (h)=\frac{1}{n}\sum_{i=1}^{n}\frac{(l_{y_i}-\hat{l}_{y_i})^2+(c_{y_i}-\hat{c}_{y_i})^2+(u_{y_i}-\hat{u}_{y_i})^2}{(1-\frac{1}{n}trL)^2} \quad (6)$$

In this paper for nonparametric fuzzy regression model taking *p=1* and *p=3*, the fuzzy local linear and fuzzy local cubic regression models are investigated with help of appropriate Taylor expansion of functions.

III. Simulation Study

By repeating the simulation experiments 200 times, the two datasets were generated which have different dimensions for selecting appropriate smoothing parameter, using same functions with Cheng and Lee`s [9] study. The fuzzy response outputs and center, right-left spreads are triangular fuzzy numbers that generated by Sample-1 and Sample-2. In simulation study, noise obtained from uniform distribution on interval [-0.5, 0.5] for center and on interval [-0.25, 0.25] for symmetric spread. In this study, it is examined that fuzzy local linear and fuzzy local cubic models are calculated for generated data sets by helping of these functions.

Sample-1. Let $g_1(x)$ is a function that defined on interval [0,10] :

$$g_1(x)=\frac{1}{5}x^2+2\exp\left(\frac{x}{10}\right)$$

For $x_i=0.1i(i=1,2,...,100)$ points on interval [0,10]

$$y_i=g_1(x_i)+rand[-0.5,0.5]$$
$$\sigma_i=\frac{1}{4}g_1(x_i)+rand[-0.25,0.25]$$

are calculated. Here $rand[a_1,a_2]$ denotes a random number generated from the uniform distribution on interval $[a_1,a_2]$ for each *i*. Assuming that the observed fuzzy outputs are symmetric triangular fuzzy numbers, so the presentation of them will be as following, *i=1,2,...,100.*

$$Y_i=(l_{y_i},c_{y_i},u_{y_i})_T=(y_i-\sigma_i,y_i,y_i+\sigma_i)_T$$

Sample-2. The second function defined as following on interval [0,10]

$$g_2(x_i)=10+5\sin(0.25\pi(1-x^2))$$

For same $x_i=0.1i(i=1,2,...,100)$ points on interval [0,10]

$$y_i=g_2(x_i)+rand[-0.5,0.5]$$
$$\sigma_i=\frac{1}{3}g_2(x_i)+rand[-0.25,0.25]$$

are calculated. Fuzzy outputs are defined;

$$Y_i=(l_{y_i},c_{y_i},u_{y_i})_T=(y_i-\sigma_i,y_i,y_i+\sigma_i)_T$$

In this study, we considered the relationship between the smoothing parameter value and sample size as a simulation study in nonparametric fuzzy local polynomial regression. For this aim, we developed fuzzy version of generalized cross-validation criteria (GCV) for selecting smoothing parameter in nonparametric fuzzy local polynomial models. Besides the local linear models, local cubic models are also used in these simulations. The appropriate smoothing parameters are selected by GCV criteria for different sample size and then performances of the models are compared using these appropriate smoothing parameters with sample sizes. 200 repetitions were made to reduce the impact of randomness in simulation calculations. The Gaussian kernel used to generate the weight sequence in the applications that made with both data sets.

IV. Conclusion

Bandwidth (h) selection plays a crucial role in local polynomial nonparametric regression problem and this is related with order of the local polynomial. Local bandwidth can increase while order of polynomial increases.

In this paper, we considered the relationship between the smoothing parameter value and sample size as a simulation study in nonparametric fuzzy local polynomial regression. The local linear models, local cubic models are also used in these simulations. Data are derived randomly by using specific functions for simulation study. It's developed fuzzy version of generalized cross-validation criteria (GCV) for selecting smoothing parameter in nonparametric fuzzy local polynomial models.

As a result, it is obtained that the bandwidth size decreases while sample size increases. This leads to increase of local fitting points and as a result overall operations increase. So, selecting appropriate bandwidth and sample size are important issue for nonparametric fuzzy local polynomial regression models.

References

[1] P. Diamond (1988), Fuzzy Least Squares, Information Sciences 46, 141-157.

[2] C.-B. Cheng, E. S. Lee (2001), Fuzzy Regression with Radial Basis Function Networks, Fuzzy Sets and Systems 119, 291-301.

[3] H. Ishibuchi, H. Tanaka (1992), Fuzzy Regression Analysis Using Neural Networks, Fuzzy Sets and Systems 50, 257-265.

[4] J. Fan and I. Gijbels (1996), Local Polynomial Modeling and Its Applications, Chapman & Hall/CRC.

[5] C. Loader (1999), Local Regression and Likelihood. Springer

[6] W. Hardle (1990) Applied Nonparametric Regression, Cambridge University Press, New York.

[7] T. Hastie, R. Tibshirani (1990), Generalized Additive Models, Chapman&Hall, London.

[8] S. N. Wood (2006), Generalized Additive Models (An Introduction with R), Chapman&Hall/CRC.

[9] C.-B. Cheng, E.S. Lee (1999), Nonparametric Fuzzy Regression – k-NN and Kernel Smoothing Techniques, Computers and Mathematical with Applications 38, 239-251.

[10] N. Wang, W.-X. Zhang, C.-L. Mei (2007), Fuzzy Nonparametric Regression Based on Local Linear Smoothing Technique, An International Journal Information Sciences, 177, 3882-3900.