COMPUTATIONAL & APPLIED MATHEMATICS

Volume 31, N. 2, pp. 323–338, 2012

ISSN 0101-8205 / ISSN 1807-0302 (Online)
www.scielo.br/cam

Integrating Ridge-type regularization in fuzzy nonlinear regression

R. FARNOOSH[1*], J. GHASEMIAN[1] and O. SOLAYMANI FARD[2]

[1]School of Mathematics, Iran University of Science and Technology, Narmak, Tehran 16844, Iran

[2]School of Mathematics and Computer Science, Damghan University, Damghan, Iran

E-mails: rfarnoosh@iust.ac.ir / jghasemian@iust.ac.ir / osfard@du.ac.ir

Abstract. In this paper, we deal with the ridge-type estimator for fuzzy nonlinear regression models using fuzzy numbers and Gaussian basis functions. Shrinkage regularization methods are used in linear and nonlinear regression models to yield consistent estimators. Here, we propose a weighted ridge penalty on a fuzzy nonlinear regression model, then select the number of basis functions and smoothing parameter. In order to select tuning parameters in the regularization method, we use the Hausdorff distance for fuzzy numbers which was first suggested by Dubois and Prade [8]. The cross-validation procedure for selecting the optimal value of the smoothing parameter and the number of basis functions are fuzzified to fit the presented model. The simulation results show that our fuzzy nonlinear modelling performs well in various situations.

Mathematical subject classification: Primary: 62J86; Secondary: 62J07.

Key words: fuzzy nonlinear regression, regularization method, Monte Carlo method, Gaussian basis expansion.

1 Introduction

Finding the relationships if any, existing in a set of variables when at least one is random, is known as an important task in statistics. On one hand, regression analysis, especially nonlinear regression, is an essential tool to analyze data.

#CAM-371/11. Received: 16/V/11. Accepted: 21/X/11.
*Corresponding author.

Many researchers use nonlinear regression more than any other statistical tool. Nonlinear models have been applied to a wide range of situations, even to finite populations. These models tend to be used either when they are suggested by theoretical considerations or to build known nonlinear behavior into a model (see for example [22, 27, 30]).

On the other hand, in most statistical practices, particularly in biology, business or government, the underlying processes are generally complex and not well understood. Also, a model obtained as the solution of a differential equation generating from engineering, chemistry, or physics is usually nonlinear. Much applied work using linear models represents a distortion of the underlying subject matter. This means that we have no idea about the form of the relationship (see [12, 27]). One of the major advantages in using nonlinear regression is the wide range of functions that can be fit.

In many practical situations, it may be unrealistic to predetermine a fuzzy parametric regression especially for a large data set with a complicated underlying variation trend. In this respect, some other approaches have been developed to deal with the fuzzy regression problems without predefining a specific form of the underlying regression relationship or with a nonlinear form of regression relationship.

The multicolinearity [21] among the independent variables leads to increasing error in estimating of regression coefficients. Shrinkage estimators have been developed for many situations including the linear and nonlinear regression models. To yield consistent estimators, the nonlinear models, instrumental variables estimation seems necessary. For this reason, the use of shrinkage estimation to nonlinear settings is recommended. The Poisson regression model [25] and the probit model [1] as well as to the Box-Cox transformation [18] are cases of such models. These papers provide theoretical results indicating the superior performance in terms of risk of certain shrinkage estimators over unrestricted estimation [10].

One of the most commonly used regularization methods is ridge regularization, which was first used in the context of least square regression in [17]. We consider applying this regularization method to nonlinear regression models involving basis expansions, which are useful tools to analyze the data with complex

structures. Expressing regression functions as a linear combination of known nonlinear functions called basis functions is the main purpose in basis expansions. However, natural cubic splines, B-splines, Fourier series and radial basis functions are widely used as basis functions [24]. In the current study, we use Gaussian basis functions [4] because they can be expressed in a simple form and can be easily implemented. When Gaussian basis functions are constructed for unequally spaced data, narrow basis functions with very small dispersions might be constructed, which can unsmoothing or unstable results if we employ these bases. Therefore, to overcome this problem, a shrinkage estimation, which allows to avoid these effects on the narrow basis functions by estimating their coefficients towards exactly zero can be used [29].

The structure of paper is as follows: Section 2 explains basic concepts of fuzzy numbers which we used in this paper. In Section 3, multivariate fuzzy nonlinear regression model based on ridge estimation will be presented. In Section 4, we discuss the selection of number basis functions and the smoothing parameter. Finally, in the last two sections, some numerical examples and comments are given.

2 Preliminaries

In this section, we read some definitions and introduce the notations which will be used throughout the paper.

Definition 2.1. ([3]) *Let X be a nonempty set. A fuzzy set $\tilde{u}$ in X is characterized by its membership function $\tilde{u} : X \to [0, 1]$. For each $x \in X$, $\tilde{u}(x)$ is interpreted as the degree of membership of an element x in the fuzzy set $\tilde{u}$.*

Let us denote by R_F the class of fuzzy subsets of the real axis R (*i.e.* $\tilde{u} : R \to [0, 1]$) satisfying the following properties:

(i) $\tilde{u}$ is normal, *i.e.*, there exists $s_0 \in R$ such that $\tilde{u}(s_0) = 1$,

(ii) $\tilde{u}$ is a convex fuzzy set, *i.e.,*

$$\tilde{u}(ts + (1 - t)r) \geq \min\{\tilde{u}(s), \tilde{u}(r)\}, \ \forall t \in [0, 1], \ s, r \in R$$

(iii) $\tilde{u}$ is upper semi-continuous on R,

(iv) $cl\{s \in R \mid \tilde{u}(s) > 0\}$ is compact, where cl denotes the closure of a subset.

R_F is called the space of fuzzy numbers, and obviously $R \subset R_F$.

For $0 < \alpha \leq 1$ denote $[\tilde{u}]^{\alpha} = \{s \in R \mid \tilde{u}(s) \geq \alpha\}$ and $[\tilde{u}]^{0} = cl\{s \in R \mid \tilde{u}(s) > 0\}$. It is clear that the $\alpha-$level set of a fuzzy number is a closed and bounded interval $[\underline{u}_{\alpha}, \overline{u}_{\alpha}]$, where $\underline{u}_{\alpha}$ denotes the left-hand endpoint of $[\tilde{u}]^{\alpha}$ and $\overline{u}_{\alpha}$ denotes the right-hand endpoint of $[\tilde{u}]^{\alpha}$.

Another definition for a fuzzy number is as follows.

Definition 2.2. ([3]) *A fuzzy number $\tilde{u}$ in parametric form is a pair $(\underline{u}, \overline{u})$ of functions $\underline{u}(\alpha)$, and $\overline{u}(\alpha)$, $0 \leq \alpha \leq 1$, which satisfies the following requirements:*

(i) $\underline{u}(\alpha)$ *is a bounded non-decreasing left continuous function in* $(0, 1]$*, and right continuous at* 0*,*

(ii) $\overline{u}(\alpha)$ *is a bounded non-increasing left continuous function in* $(0, 1]$*, and right continuous at* 0*,*

(iii) $\underline{u}(\alpha) \leq \overline{u}(\alpha)$*,* $0 \leq \alpha \leq 1$.

A crisp number a is simply represented by $\underline{u}(\alpha) = \overline{u}(\alpha) = a, 0 \leq \alpha \leq 1$. We recall that for $a < b < c$, which $a, b, c \in R$, the triangular fuzzy number $\tilde{u} = (a, b, c)$ determined by a, b, c is given such that $\underline{u}(\alpha) = a + (b - c)\alpha$ and $\bar{u}(\alpha) = c - (c - b)\alpha$ are the endpoints of the α-level sets, for all $\alpha \in [0, 1]$.

For arbitrary $\tilde{u} = (\underline{u}(\alpha), \overline{u}(\alpha))$, $\tilde{v} = (\underline{v}(\alpha), \overline{v}(\alpha))$, we define the following

(i) $(\tilde{u} \oplus \tilde{v})(\alpha) = (\underline{u}(\alpha) + \underline{v}(\alpha), \overline{u}(\alpha) + \overline{v}(\alpha)), \forall \alpha \in [0, 1]$,

(ii) $(\tilde{u} - \tilde{v})(\alpha) = (\underline{u}(\alpha) - \overline{v}(\alpha), \overline{u}(\alpha) - \underline{v}(\alpha)), \forall \alpha \in [0, 1]$,

(iii) $(k \otimes \tilde{u})(\alpha) = \begin{cases} (k\underline{u}(\alpha), k\overline{u}(\alpha)), k \geq 0 \\ (k\overline{u}(\alpha), k\underline{u}(\alpha)), k < 0 \end{cases}, \forall \alpha \in [0, 1]$,

where k is a scalar. Moreover, when $k = -1$, we have $k \otimes \tilde{u} = -\tilde{u}$.

The Hausdorff distance between fuzzy numbers given by $D : R_F \times R_F \rightarrow R_+ \bigcup \{0\}$,

$$D(\tilde{u}, \tilde{v}) = \sup_{\alpha \in [0,1]} \max \left\{ \left| \underline{u}(\alpha) - \underline{v}(\alpha) \right|, \left| \overline{u}(\alpha) - \overline{v}(\alpha) \right| \right\},$$

where, $\tilde{u} = (\underline{u}(\alpha), \overline{u}(\alpha))$, $\tilde{v} = (\underline{v}(\alpha), \overline{v}(\alpha)) \subset R$.

Note that (R_F, D) is a complete metric space (see [3, 11, 13]) and has the following properties,

(i) $D(\tilde{u} \oplus \tilde{w}, \tilde{v} \oplus \tilde{w}) = D(\tilde{u}, \tilde{v}), \forall\, \tilde{u}, \tilde{v}, \tilde{w} \in R_F$,

(ii) $D(k \otimes \tilde{u}, k \otimes \tilde{v}) = |k|\, D(\tilde{u}, \tilde{v}), \forall\, k \in R, \tilde{u}, \tilde{v} \in R_F$,

(iii) $D(\tilde{u} \oplus \tilde{v}, \tilde{w} \oplus \tilde{e}) \leq D(\tilde{u}, \tilde{w}) + D(\tilde{v}, \tilde{e}), \forall\, \tilde{u}, \tilde{v}, \tilde{w}, \tilde{e} \in R_F$.

Definition 2.3. ([3]) *A mapping $f : R_F \to R_F$ is called a fuzzy process. Therefore, its α-level set can be written as follows,*

$$[f(\tilde{x})]^\alpha = [\underline{f}^\alpha(\tilde{x}), \overline{f}^\alpha(\tilde{x})], \tilde{x} \in R_F, \alpha \in [0, 1].$$

Definition 2.4. ([11]) *A mapping $f : R_F \to R_F$ is called continuous at point $\tilde{x}_0 \in R_F$ provided for any fixed $\alpha \in [0, 1]$ and arbitrary $\varepsilon > 0$, there exists $\delta(\varepsilon, \alpha)$ such that $D([f(\tilde{x})]^\alpha, [f(\tilde{x}_0)]^\alpha) < \varepsilon$, whenever $D([\tilde{x}]^\alpha, [\tilde{x}_0]^\alpha) < \delta(\varepsilon, \alpha)$ for all $\tilde{x} \in R_F$.*

3 Multivariate Fuzzy Nonlinear Regression Model Based on Ridge Estimation

A fuzzy nonlinear regression model with multivariate fuzzy input and output is considered in this section and, a fitting procedure is proposed for this model.

Suppose that we have n independent observations $(\tilde{y}_i, \tilde{\mathbf{x}}_i); i = 1, 2, \ldots, n$. Here, a fuzzy nonlinear regression model is considered as follows,

$$\tilde{y}_i = m(\tilde{\mathbf{x}}_i) \oplus \tilde{\varepsilon}_i, i = 1, 2, \ldots, n \tag{1}$$

In this model, $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip}) \in (R_F)^p;\ i = 1, 2, \ldots, n$, are vectors of p-dimensional independent variables(inputs) and $\tilde{y}_i \in R_F$ is univariate output. The function $m(.)$, a mapping from $(R_F)^p$ to R_F, is an unknown fuzzy smooth function. Moreover, $\tilde{\varepsilon}_i \in R_F$ for which $\tilde{\varepsilon}_i^\alpha = (\underline{\varepsilon}_i(\alpha), \overline{\varepsilon}_i(\alpha))$, and $\underline{\varepsilon}_i(\alpha)$ and $\overline{\varepsilon}_i(\alpha))$ are independently, normally distributed with mean zero and variance $\sigma^2(\alpha)$. Regarding Definition 2.3, the fuzzy function $m(.)$ can be expressed as

$$m^\alpha(\tilde{x}_{ij}) = \left[\underline{m}^\alpha(x_{ij}), \overline{m}^\alpha(x_{ij})\right]; i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p.$$

So, we can write the model $\tilde{y} = m(\tilde{\mathbf{x}}) \oplus \tilde{\varepsilon}$, as the follow form,

$$\begin{cases} \underline{y}(\alpha) = \underline{m}^{\alpha}(\tilde{\mathbf{x}}) + \underline{\varepsilon}(\alpha) \\ \overline{y}(\alpha) = \overline{m}^{\alpha}(\tilde{\mathbf{x}}) + \overline{\varepsilon}(\alpha) \end{cases} \tag{2}$$

that equivalent to

$$\begin{cases} \underline{y}(\alpha) = \underline{m}(\underline{\mathbf{x}}, \overline{\mathbf{x}}; \alpha) + \underline{\varepsilon}(\alpha) \\ \overline{y}(\alpha) = \overline{m}(\underline{\mathbf{x}}, \overline{\mathbf{x}}; \alpha) + \overline{\varepsilon}(\alpha) \end{cases} \tag{3}$$

We may assume any component of $\underline{m}(.)$ or $\overline{m}(.)$ has a linear combination of basis functions $\phi_l(\underline{\mathbf{x}}, \overline{\mathbf{x}}, \alpha)$; $l = 1, \ldots, k$ in the form

$$\underline{m}(\underline{\mathbf{x}}, \overline{\mathbf{x}}, \alpha) = \underline{b}_0(\alpha) + \sum_{l=1}^{k} \underline{b}_l(\alpha)\phi_l(\underline{\mathbf{x}}, \overline{\mathbf{x}}, \alpha) \tag{4}$$

and

$$\overline{m}(\underline{\mathbf{x}}, \overline{\mathbf{x}}, \alpha) = \overline{b}_0(\alpha) + \sum_{l=1}^{k} \overline{b}_l(\alpha)\phi_l(\underline{\mathbf{x}}, \overline{\mathbf{x}}, \alpha) \tag{5}$$

where

$$\underline{\mathbf{b}}(\alpha) = (\underline{b}_0(\alpha), \underline{b}_1(\alpha), \ldots, \underline{b}_k(\alpha))^T \text{ and } \overline{\mathbf{b}}(\alpha) = (\overline{b}_0(\alpha), \overline{b}_1(\alpha), \ldots, \overline{b}_k(\alpha))^T$$

are fuzzy unknown coefficient parameter vectors and for $\mathbf{X}^{\alpha} = (\underline{\mathbf{x}}, \overline{\mathbf{x}}, \alpha)$, Gaussian basis functions are given by

$$\begin{cases} \phi_l(\mathbf{X}^{\alpha}, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha)) = \exp\left(-\frac{||\mathbf{X}^{\alpha} - \underline{\boldsymbol{\mu}}_l(\alpha)||^2}{2\underline{h}_l^2(\alpha)}\right), \\ \phi_l(\mathbf{X}^{\alpha}, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha)) = \exp\left(-\frac{||\mathbf{X}^{\alpha} - \overline{\boldsymbol{\mu}}_l(\alpha)||^2}{2\overline{h}_l^2(\alpha)}\right), \end{cases} \quad (l = 1, \ldots, k) \tag{6}$$

where $\underline{\boldsymbol{\mu}}_l(\alpha)$ and $\overline{\boldsymbol{\mu}}_l(\alpha)$ are p-dimensional vector determining the center of the basis functions, $\underline{h_l}^2(\alpha)$ and $\overline{h_l}^2(\alpha)$ are the width parameters and $||.||$ is the Euclidian norm. Unknown parameters in the models (4) and (5) include the coefficient parameters $\mathbf{b}(\alpha) = (\underline{\mathbf{b}}(\alpha), \overline{\mathbf{b}}(\alpha))$, and the centers $\underline{\boldsymbol{\mu}}_l(\alpha)$, $\overline{\boldsymbol{\mu}}_l(\alpha)$ and width parameters $\underline{h_l}^2(\alpha)$, $\overline{h_l}^2(\alpha)$ required for Gaussian basis functions. These parameters are generally determined in a two-stage procedure in order to avoid local minimum

and identification problems. In the first stage, the centers $\underline{\boldsymbol{\mu}}_l(\alpha)$, $\overline{\boldsymbol{\mu}}_l(\alpha)$ and dispersion $\underline{h_l}^2(\alpha)$, $\overline{h_l}^2(\alpha)$ are determined by using the k-means clustering algorithm. The data set of observations of the explanatory variables $(\underline{\mathbf{x}}_1(\alpha), \ldots, \underline{\mathbf{x}}_n(\alpha))$ and $(\overline{\mathbf{x}}_1(\alpha), \ldots, \overline{\mathbf{x}}_n(\alpha))$ are divided respectively into k clusters $(\underline{\mathbf{C}}_1(\alpha), \ldots, \underline{\mathbf{C}}_k(\alpha))$ and $(\overline{\mathbf{C}}_1(\alpha), \ldots, \overline{\mathbf{C}}_k(\alpha))$; centers $\underline{\boldsymbol{\mu}}_l(\alpha)$, $\overline{\boldsymbol{\mu}}_l(\alpha)$ and dispersions $\underline{h_l}^2(\alpha)$, $\overline{h_l}^2(\alpha)$ are determined by

$$\begin{cases} \hat{\underline{\boldsymbol{\mu}}}_l(\alpha) = \frac{1}{n_l} \sum_{\underline{\mathbf{x}}_i(\alpha) \in \underline{\mathbf{C}}_l} \underline{\mathbf{x}}_i(\alpha), \\ \hat{\overline{\boldsymbol{\mu}}}_l(\alpha) = \frac{1}{n_l} \sum_{\overline{\mathbf{x}}_i(\alpha) \in \overline{\mathbf{C}}_l} \overline{\mathbf{x}}_i(\alpha), \\ \hat{\underline{h}_l^2}(\alpha) = \frac{1}{n_l} \sum_{\underline{\mathbf{x}}_i(\alpha) \in \underline{\mathbf{C}}_l} ||\underline{\mathbf{x}}_i(\alpha) - \underline{\mu}_l(\alpha)||^2, \\ \hat{\overline{h}_l^2}(\alpha) = \frac{1}{n_l} \sum_{\overline{\mathbf{x}}_i(\alpha) \in \overline{\mathbf{C}}_l} ||\overline{\mathbf{x}}_i(\alpha) - \overline{\boldsymbol{\mu}}_l(\alpha)||^2, \end{cases} \quad (l = 1, \ldots, k) \tag{7}$$

where n_l is the number of observations included in the l-th cluster $\underline{\mathbf{C}}_l$ or $\overline{\mathbf{C}}_l$. Replacing $\mu_l(\alpha) = (\underline{\boldsymbol{\mu}}_l(\alpha), \overline{\boldsymbol{\mu}}_l(\alpha))$ and $h_l^2(\alpha) = (\underline{h_l}^2(\alpha), \overline{h_l}^2(\alpha))$ in (6) by (7) respectively, we obtain a set of $2k$ basis functions

$$\begin{cases} \phi_l(\mathbf{X}^\alpha, \hat{\underline{\boldsymbol{\mu}}}_l(\alpha), \hat{\underline{h}}_l(\alpha)) = \exp\left(-\frac{||\mathbf{X}^\alpha - \hat{\underline{\boldsymbol{\mu}}}_l(\alpha)||^2}{2\hat{\underline{h}}_l^2(\alpha)}\right), \\ \phi_l(\mathbf{X}^\alpha, \hat{\overline{\boldsymbol{\mu}}}_l(\alpha), \hat{\overline{h}}_l(\alpha)) = \exp\left(-\frac{||\mathbf{X}^\alpha - \hat{\overline{\boldsymbol{\mu}}}_l(\alpha)||^2}{2\hat{\overline{h}}_l^2(\alpha)}\right), \end{cases} \quad (l = 1, \ldots, k) \tag{8}$$

For n independent observations $(\tilde{y}_i, \tilde{\mathbf{x}}_i)$; $i = 1, 2, \ldots, n$, the fuzzy nonlinear regression model based on Gaussian basis functions

$$\phi_l(\mathbf{X}^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha)), \phi_l(\mathbf{X}^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha)), l = 1, \ldots, k$$

given in (6) is expressed as

$$\begin{cases} \underline{y}_i(\alpha) = \underline{\mathbf{b}}^T(\alpha)\phi(\mathbf{X}_i^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha)) + \underline{\varepsilon}_i(\alpha) \\ \overline{y}_i(\alpha) = \overline{\mathbf{b}}^T(\alpha)\phi(\mathbf{X}_i^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha)) + \overline{\varepsilon}_i(\alpha) \end{cases} \tag{9}$$

where

$$\boldsymbol{\phi}(\mathbf{X}_i^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha)) = (1, \phi_1(\mathbf{X}_i^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha)), \ldots, \phi_k(\mathbf{X}_i^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha))),$$

$$\boldsymbol{\phi}(\mathbf{X}_i^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha)) = (1, \phi_1(\mathbf{X}_i^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha)), \ldots, \phi_k(\mathbf{X}_i^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha))),$$

$$\underline{\mathbf{b}}(\alpha) = (\underline{b}_0(\alpha), \underline{b}_1(\alpha), \ldots, \underline{b}_k(\alpha))^T, \overline{\mathbf{b}}(\alpha) = (\overline{b}_0(\alpha), \overline{b}_1(\alpha), \ldots, \overline{b}_k(\alpha))^T,$$

and $\underline{\varepsilon}_i(\alpha), \overline{\varepsilon}_i(\alpha)$ are error terms. If the error terms are independently and normally distributed with mean 0 and variance $h_l^2(\alpha) = (\underline{h_l}^2(\alpha), \overline{h_l}^2(\alpha))$, for all $\alpha \in [0, 1]$ the nonlinear regression model (9) has a probability density function

$$\begin{cases} f\left(\underline{y}_i(\alpha)|\mathbf{X}_i^\alpha; \underline{\mathbf{b}}(\alpha), \underline{h}^2(\alpha)\right) \\ \quad = \frac{1}{\sqrt{2\pi\underline{h}^2(\alpha)}} \exp\left[-\frac{\{\underline{y}_i(\alpha)-\underline{\mathbf{b}}^T(\alpha)\boldsymbol{\phi}(\mathbf{X}_i^\alpha)\}^2}{2\hat{\underline{h}}^2(\alpha)}\right], \\ f\left(\overline{y}_i(\alpha)|\mathbf{X}_i^\alpha; \overline{\mathbf{b}}(\alpha), \overline{h}^2(\alpha)\right) \\ \quad = \frac{1}{\sqrt{2\pi\overline{h}^2(\alpha)}} \exp\left[-\frac{\{\overline{y}_i(\alpha)-\overline{\mathbf{b}}^T(\alpha)\boldsymbol{\phi}(\mathbf{X}_i^\alpha)\}^2}{2\hat{\overline{h}}^2(\alpha)}\right], \quad (i = 1, \ldots, n) \end{cases} \tag{10}$$

Then the maximum likelihood estimates of the coefficient vectors $\underline{\mathbf{b}}(\alpha), \overline{\mathbf{b}}(\alpha)$ and $h_l^2(\alpha) = (\underline{h_l}^2(\alpha), \overline{h_l}^2(\alpha))$ are respectively given by

$$\begin{cases} \hat{\underline{\mathbf{b}}}(\alpha) = (\underline{\Phi}^T\underline{\Phi})^{-1}\underline{\Phi}^T\underline{\mathbf{y}}(\alpha) \\ \hat{\overline{\mathbf{b}}}(\alpha) = (\overline{\Phi}^T\overline{\Phi})^{-1}\overline{\Phi}^T\overline{\mathbf{y}}(\alpha) \\ \hat{\underline{h}}^2(\alpha) = \frac{1}{n}[\underline{\mathbf{y}}(\alpha) - \underline{\Phi}\,\hat{\underline{\mathbf{b}}}(\alpha)]^T[\underline{\mathbf{y}}(\alpha) - \underline{\Phi}\,\hat{\underline{\mathbf{b}}}(\alpha)] \\ \frac{1}{n}[\overline{\mathbf{y}}(\alpha) - \overline{\Phi}\,\hat{\overline{\mathbf{b}}}(\alpha)]^T[\overline{\mathbf{y}}(\alpha) - \overline{\Phi}\,\hat{\overline{\mathbf{b}}}(\alpha)] \end{cases} \tag{11}$$

where

$$\underline{\Phi} = (\phi(\mathbf{X}_1^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha)), \ldots, \phi(\mathbf{X}_n^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha))),$$

$$\overline{\Phi} = (\phi(\mathbf{X}_1^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha)), \ldots, \phi(\mathbf{X}_n^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha))),$$

$$\underline{\mathbf{y}}(\alpha) = (\underline{y}_1(\alpha), \underline{y}_2(\alpha), \ldots, \underline{y}_n(\alpha))^T,$$

$$\overline{\mathbf{y}}(\alpha) = (\overline{y}_1(\alpha), \overline{y}_2(\alpha), \ldots, \overline{y}_n(\alpha))^T.$$

We estimate $\mathbf{b}(\alpha) = (\underline{\mathbf{b}}(\alpha), \overline{\mathbf{b}}(\alpha))$ and $h^2(\alpha)$ by the regularization method, because the maximum likelihood method often yields unstable estimates in fitting a nonlinear model to data with a complex structure. We consider maximizing the penalized log-likelihood function, instead of using the log-likelihood function,

$$\begin{cases} l_\lambda(\underline{\boldsymbol{\theta}}(\alpha)) = \sum_{i=1}^n \log f\left(\underline{y}_i^\alpha|\mathbf{X}_i^\alpha; \underline{\mathbf{b}}(\alpha), \underline{h}(\alpha)\right) - n\lambda H(\underline{\mathbf{b}}(\alpha)) \\ l_\lambda(\overline{\theta}(\alpha)) = \sum_{i=1}^n \log f\left(\overline{y}_i^\alpha|\mathbf{X}_i^\alpha; \overline{\mathbf{b}}(\alpha), \overline{h}(\alpha)\right) - n\lambda H(\overline{\mathbf{b}}(\alpha)) \end{cases} \tag{12}$$

where

$$\underline{\theta}(\alpha) = (\underline{\mathbf{b}}^T(\alpha), \underline{h}^2(\alpha))^T, \overline{\theta}(\alpha) = (\overline{\mathbf{b}}^T(\alpha), \overline{h}^2(\alpha))^T,$$

and $H(\mathbf{b}(\alpha)) = [H(\underline{\mathbf{b}}(\alpha)), H(\overline{\mathbf{b}}(\alpha))]$ is a penalty function for $\mathbf{b}(\alpha)$ and $\lambda(> 0)$ is a smoothing parameter that controls the smoothness of the fitted model. Based on the ridge penalty, in l^2 norm, $(H(\underline{\mathbf{b}}(\alpha)), H(\overline{\mathbf{b}}(\alpha)))$ given by

$$\begin{cases} H(\underline{\mathbf{b}}(\alpha)) = \frac{1}{2} \sum_{l=1}^{k} \underline{b}_l{}^2(\alpha) = \frac{1}{2}\underline{\mathbf{b}}^T(\alpha)\underline{\mathbf{b}}(\alpha), \\ H(\overline{\mathbf{b}}(\alpha)) = \frac{1}{2} \sum_{l=1}^{k} \overline{b}_l{}^2(\alpha) = \frac{1}{2}\overline{\mathbf{b}}^T(\alpha)\overline{\mathbf{b}}(\alpha). \end{cases} \tag{13}$$

Then, the maximum penalized likelihood estimates of $\mathbf{b}(\alpha) = (\underline{\mathbf{b}}(\alpha), \overline{\mathbf{b}}(\alpha))$ and $h^2(\alpha) = (\underline{h}^2(\alpha), \overline{h}^2(\alpha))$ are respectively given by

$$\begin{cases} \hat{\underline{\mathbf{b}}}(\alpha) = [\underline{\Phi}^T\underline{\Phi} + n\lambda\hat{\underline{h}}^2(\alpha) I_{l+1}]^{-1}\underline{\Phi}^T\underline{\mathbf{y}}(\alpha), \\ \hat{\overline{\mathbf{b}}}(\alpha) = [\overline{\Phi}^T\overline{\Phi} + n\lambda\hat{\overline{h}}^2(\alpha) I_{l+1}]^{-1}\overline{\Phi}^T\overline{\mathbf{y}}(\alpha), \\ \hat{\underline{h}}^2(\alpha) = \frac{1}{n}[\underline{\mathbf{y}}(\alpha) - \underline{\Phi}\,\hat{\underline{\mathbf{b}}}(\alpha)]^T[\underline{\mathbf{y}}(\alpha) - \underline{\Phi}\,\hat{\underline{\mathbf{b}}}(\alpha)], \\ \hat{\overline{h}}^2(\alpha) = \frac{1}{n}[\overline{\mathbf{y}}(\alpha) - \overline{\Phi}\,\hat{\overline{\mathbf{b}}}(\alpha)]^T[\overline{\mathbf{y}}(\alpha) - \overline{\Phi}\,\hat{\overline{\mathbf{b}}}(\alpha)]. \end{cases} \tag{14}$$

I_{l+1} is an $(l+1)$ dimensional identity matrix. Note that these estimators depend on each other. Therefore, we provide an initial value for the variance $\left(\underline{h}^2{}_{\tilde{\mathbf{X}}^\alpha(0)}(\alpha), \overline{h}^2_{\tilde{\mathbf{X}}^\alpha(0)}(\alpha)\right)$ first, then $(\hat{\underline{\mathbf{b}}}(\alpha), \hat{\overline{\mathbf{b}}}(\alpha))$ and $\left(\underline{h}^2{}_{\tilde{\mathbf{X}}^\alpha}(\alpha), \overline{h}^2_{\tilde{\mathbf{X}}^\alpha}(\alpha)\right)$ are updated until convergence. This estimation method is the maximum penalized likelihood with the quadratic form (see [9, 23]). The consistency and the rate of convergence of these estimators proved in ([5, 15, 20]).

4 Selection of the number of basis function and the smoothing parameter

The estimates that achieved in (14) by the regularization method depends upon the number of basis functions $2k$ and the value of the smoothing parameter λ. Appropriate determining these values is a crucial issue. The fuzzified cross-validation procedure based on the Hausdorff distance between fuzzy numbers, can be described as follows. Let

$$\begin{cases} \underline{\hat{y}}_i(\alpha) = \underline{\hat{m}}^\alpha(\tilde{\mathbf{x}}_i) = \hat{\underline{\mathbf{b}}}^T(\alpha)\boldsymbol{\phi}(\mathbf{X}_i^\alpha, \underline{\boldsymbol{\mu}}_l(\alpha), \underline{h}_l(\alpha)) \\ \overline{\hat{y}}_i(\alpha) = \hat{\overline{m}}^\alpha(\tilde{\mathbf{x}}_i) = \hat{\overline{\mathbf{b}}}^T(\alpha)\boldsymbol{\phi}(\mathbf{X}_i^\alpha, \overline{\boldsymbol{\mu}}_l(\alpha), \overline{h}_l(\alpha)), \end{cases} \tag{15}$$

be the predicted fuzzy ridge nonlinear regression function at input $\mathbf{X}_i^{\alpha}$ computed by our method. Let

$$CV(k,\lambda)=\frac{1}{n}\sum_{i=1}^{n} D(\tilde{y}_i,\hat{\tilde{y}}_i). \tag{16}$$

The $CV(k,\lambda)$ quantity gives an overall measurement of the difference between the actual values of dependent variable and its estimation. However, because of the error term in model (1) the $CV(k,\lambda)$ cannot efficiently reflect the closeness between the underlying fuzzy nonlinear regression function $m(\tilde{\mathbf{x}})$ and its estimate. Therefore, we define a quantity for measuring the bias between the objective function and its estimate, which is

$$BIAS(k,\lambda)=\frac{1}{n}\sum_{i=1}^{n} D(m(\tilde{\mathbf{x}}_i),\hat{m}(\tilde{\mathbf{x}}_i)). \tag{17}$$

The $BIAS(k,\lambda)$ makes sense for examining the performance of the different methods by simulation. Both $CV(k,\lambda)$ and $BIAS(k,\lambda)$ will be reported in our simulations to numerically evaluate the performance of the proposed method. Choose k_0 and λ_0 as the optimal values such that

$$CV(k_0,\lambda_0) = \min_{k>0,\lambda>0} CV(k,\lambda), \quad \text{and,}$$

$$BIAS(k_0,\lambda_0) = \min_{k>0,\lambda>0} BIAS(k,\lambda).$$

In practice, we may compute for a series of values of k and λ to obtain k_0 and λ_0. A smoother regression function generally corresponds to a larger value of k while a more fluctuating regression function tends to select a smaller value of k.

5 Numerical results

In this section, we use Monte Carlo simulations to investigate the performance of the proposed model by computing $CV(k,\lambda)$ and $BIAS(k,\lambda)$. Here, two simulation examples are considered: a curve fitting and a surface fitting. The results are obtained for some values of sample size, k and λ.

Example 5.1. *Curve fitting.* In this simulation repeated random samples $(\tilde{y}_i,\tilde{x}_i)$; $i = 1,2,\ldots,n$ with $n = 100$ or 120 were generated from a true nonlinear

regression model $\tilde{y}_i = m(\tilde{x}_i) \oplus \tilde{\varepsilon}_i$ where

$$m(\tilde{x}) = \exp(2\tilde{x}^2).$$

The design points $\tilde{x}_i$ are uniformly distributed in $[0, 1]_F \subseteq R_F$ and the errors $\tilde{\varepsilon}_i^{\alpha} = (\underline{\varepsilon}_i(\alpha), \overline{\varepsilon}_i(\alpha))$, which $\underline{\varepsilon}_i(\alpha)$ and $\overline{\varepsilon}_i(\alpha)$ are independently, normally distributed. The mean of distribution is zero and the standard deviations are respectively $\underline{\tau} = 0.1R_{\underline{m}}$ and $\overline{\tau} = 0.1R_{\overline{m}}$ with $R_{\underline{m}}$, $R_{\overline{m}}$ being the ranges of $\underline{m}(\tilde{x})$ or $\overline{m}(\tilde{x})$ over $\tilde{x}_i \in [0, 1]_F$.

$CV(k, \lambda)$ and $BIAS(k, \lambda)$ are used to numerically evaluate the performance of method and the related results are summarized in Table 1. In this example it is clear that the ridge estimation must be used.

Sample Size(n)	λ	k	$BIAS$	CV
100	0 (without ridge)	20	21.310	21.28
100	1.15e-1	50	3.99e-4	0.008
100	1.15e-5	20	4.31e-4	0.0084
100	7.15e-5	20	0.0017	0.0081
100	1.15e-2	20	0.0102	0.0114
100	1.15e-3	20	0.0043	0.009
100	1.15e-7	20	4.96e-6	0.0081
100	8.15e-10	10	0.0041	0.0082
100	1.15e-3	10	0.09	0.0896
100	7.15e-5	5	0.38	0.37
100	8.15e-10	5	0.09	0.0917
120	1.15e-3	30	6.35e-4	0.0084
120	7.15e-5	30	3.96e-5	0.0083

Table 1 – The simulation results obtained by the proposed method.

Example 5.2. *Surface fitting.* Next, we applied the modeling strategy to surface data. We generated random samples $(\tilde{y}_i, \tilde{x_1}_i, \tilde{x_2}_i)$; $i = 1, 2, \ldots, n$ with $n = 100$ or 120 from a true model $\tilde{y}_i = m(\tilde{x_1}_i, \tilde{x_2}_i) \oplus \tilde{\varepsilon}_i$ where

$$m(\tilde{x_1}, \tilde{x_2}) = \exp\left(-2\sqrt{\tilde{x_1}^2 + \tilde{x_2}^2}\right).$$

The design points $(\tilde{x_{1i}}, \tilde{x_{2i}})$ are uniformly distributed in $[0,1]_F \times [0,1]_F \subseteq (R_F)^2$. The errors $\tilde{\varepsilon}_i^{\alpha} = (\underline{\varepsilon}_i(\alpha), \overline{\varepsilon}_i(\alpha))$, which $\underline{\varepsilon}_i(\alpha)$ and $\overline{\varepsilon}_i(\alpha)$ are independently, normally distributed. The mean of distribution is zero and the standard deviations are respectively $\underline{\tau} = 0.1R_{\underline{m}}$ and $\overline{\tau} = 0.1R_{\overline{m}}$ with $R_{\underline{m}}$, $R_{\overline{m}}$ being the ranges of $\underline{m}(\tilde{x_1}, \tilde{x_2})$ or $\overline{m}(\tilde{x_1}, \tilde{x_2})$ over $(\tilde{x_1}, \tilde{x_2}) \in [0,1]_F \times [0,1]_F$.

Table 2 shows the result for this dataset. The results show that the ridge method still produces a quite satisfactory estimate of the fuzzy nonlinear regression in the case of two-dimensional input.

Sample Size(n)	λ	k	$BIAS$	CV
120	0 (without ridge)	30	6.12	6.01
100	1.15e-1	50	0.0030	0.0090
100	1.15e-5	20	2.52e-6	0.008
100	7.15e-5	20	1.57e-5	0.0085
100	1.15e-2	20	0.0023	0.0098
100	1.15e-3	20	2.50e-4	0.0083
100	1.15e-7	20	2.88e-8	0.009
100	8.15e-10	10	3.61e-6	0.0082
100	1.15e-3	10	0.0058	0.0096
100	7.15e-5	5	0.0179	0.0182
100	8.15e-10	5	0.0165	0.0177
120	1.15e-3	30	1.68e-5	0.0086
120	7.15e-5	30	0.0014	0.0093

Table 2 – The simulation results for two-dimensional dataset obtained by the proposed method.

Example 5.3. Consider the below function from [31].

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

Let

$$\begin{cases} y_i = g(x_i) + \text{rand}[-0.5, 0.5], \\ \sigma_i = \frac{1}{3}g(x_i) + \text{rand}[-0.25, 0.25], \end{cases} \quad i = 1, 2, \ldots, 100,$$

The observed fuzzy outputs are

$$\tilde{y_i} = (\underline{y}_i, \overline{y}_i) = (y_i - \sigma_i, y_i + \sigma_i), i = 1, 2, \ldots, 100.$$

Table 3 shows the result for this dataset.

Method	Kernel	λ	k	$BIAS$	CV
The proposed method	–	0.001	20	0.0031	0.0032
	–	0.001	10	0.0318	0.0321
	–	0.01	20	0.0162	0.0178
	–	0.01	10	2.8874e-4	3.2981e-4
	–	0.03	10	0.2692	0.2944
	–	0.03	20	0.1758	0.1982
	–	0.003	20	2.2701e-5	4.3991e-5
LLS	Gauss	0.21	–	0.0520	0.2252
	Epanechnikov	0.52	–	0.0552	0.2544
KS	Gauss	0.15	–	0.0821	0.2080
	Epanechnikov	0.34	–	0.0848	0.2337

Table 3 – The simulation results obtained by different methods.

In this example LLS and KS, respectively stand for the local linear smoothing and kernel smoothing methods.

6 Conclusions

In this study, we dealt with estimating the ridge regularization in fuzzy nonlinear regression model with modelling the data with multivariate fuzzy input and output. The ridge estimation of fuzzy nonlinear regression models based on Gaussian basis function with the cross-validation procedure for selecting the optimal values of the smoothing parameter and the number of basis function was proposed. Some simulation experiments were conducted to assess the performance of the method. By computing the $CV(k, \lambda)$ and $BIAS(k, \lambda)$, we found that the proposed method performs quite well in reducing the error and producing a satisfactory estimate of the parameters of nonlinear regression function.

As demonstrated by the numerical experiments, the increasing of k, the number of basis function, increases the accuracy of model. In this case we can decrease the smoothing parameter λ. The small values of $CV(k, \lambda)$ and $BIAS(k, \lambda)$ indicates that the proposed method tends to produce estimates that are more close to their right values and gives less biased estimates of the real function.

Acknowledgements. The authors are grateful to the associate editor and two referees for their accurate reading and their helpful suggestions and remarks.

REFERENCES

[1] L.C. Adkins and R.C. Hill, *Risk characteristics of a Stein-like estimator of the probit regression model*. Economics Letters, **30** (1989), 19–26.

[2] D.M. Bates and D.G. Watts, Nonlinear Regression Analysis and its Applications. Wiley, New York (1988).

[3] B. Bede and S.G. Gal, *Generalizations of the differentiability of fuzzy umber valued functions with applications to fuzzy differential equation*. Fuzzy Sets and Systems, **151** (2005), 581–599.

[4] C.M. Bishop, Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995).

[5] D.D. Cox and F. O'Sullivan, *Asymptotic Analysis of Penalized Likelihood and Related Estimators*. Ann. Statist., **18** (1990), 1676–1695.

[6] P. Diamond, *Fuzzy least squares*. Information Sciences, **46** (1988), 141–157.

[7] N.R. Draper and H. Smith, Applied Regression Analysis. Wiley, New York (1980).

[8] D. Dubois and H. Prade, *Fuzzy Numbers: An Overview, Analysis of Fuzzy Information*. Mathematical Logic,CRC Press, Boca Raton, FL, **1** (1987), 3–39.

[9] P.P.B Eggermont and V.N. Lariccia, Maximum Penalized Likelihood Estimation, Springer (2001).

[10] S. Ejaz Ahmed and C.J. Nicol, *An application of shrinkage estimation to the nonlinear regression model*. Journal of Computational Statistics and Data Analysis (to appear).

[11] M. Friedman, M. Ma and A. Kandel, *Numerical solutions of fuzzy differential and integral equations*. Fuzzy Sets and Systems, **106** (1999), 35–48.

[12] A.R. Gallant, Nonlinear Statistical Models. Wiley, New York (1987).

[13] R. Goetschel and W. Voxman, *Topological properties of fuzzy number*. Fuzzy Sets and Systems, **10** (1983), 87–99.

[14] R. Goetschel and W. Voxman, *Elementary fuzzy calculus, Fuzzy sets and Systems*, **18** (1986), 31–43.

[15] C. Gu and C. Qiu, *Penalized Likelihood regression: A Simple Asymptotic Analysis*. Statistica Sinica, **4** (1994), 297–304.

[16] T.J. Hastie and R.J. Tibshirani, Generalized additive models, Chapman & Hall, London (1990).

[17] A.E. Hoerl and R.W. Kennard, *Ridge regression: biased estimates for nonorthogonal problems*. Technometrics, **12** (1970), 55–67.

[18] M. Kim and R.C. Hill, *Shrinkage estimation in nonlinear models. The Box-Cox transformation*. Journal of Econometrics, **66** (1995), 1–33.

[19] B. Kim and R.R. Bishu, *Evaluation of fuzzy linear regression models by comparing membership functions*. Fuzzy Sets and Systems, **100** (1998), 343–352.

[20] G.A. Latham and S. Yu, *N-Stage Splitting for Maximum Penalized Likelihood Estimation with Gaussian Data and Stationary Linear Iterative Methods*. J. Statist. Comput. and Simul., **62** (1999), 375–393.

[21] R.X. Liu, J. Kuang, Q. Gong and X.L. Hou, *Principal component regression analysis with SPSS*. Computer Methods and Programs in Biomedicine, **71** (2003), 141–147.

[22] H. Motulsky and A. Christopoulos, Fitting Modeles to Biological Data using Linear and Nonlinear Regression. GraphPad Software, Inc., San Diego (2003).

[23] T. Poggio and F. Girosi, *Networks for approximation and learning*. Proc. IEEE, **78** (1990), 1484–1487.

[24] D. Ruppert, M.P. Wand and R.J. Carrol, Semiparametric Regression, Cambridge University Press (2003).

[25] S.K. Sapra, *Pre-test estimation in Poisson regression model*. Applied Economics Letters, **10** (2003), 541–543.

[26] C. Saunders, A. Gammerman and V. Vork, *Ridge regression learning algorithm in dual variable*. Proceedings of the 15th International Conference on Machine Learning, (1998), 515–521.

[27] G.A.F. Seber and C.J. Wild, Nonlinear Regression. Wiley, New York (1989).

[28] H. Tanaka and H. Ishibuchi, *Identification of possibilistic linear systems by quadratic membership functions of fuzzy parameters*. Fuzzy Sets and Systems, **41** (1991), 145–160.

[29] S. Tateishi, H. Matsui and S. Konishi, *Nonlinear regression modeling via the lasso-type regularization*. Journal of Statistical Planning and Inference, **140** (2010), 1125–1134.

[30] R. Valiant, *Nonlinear prediction theory and the estimation of proportions in a finite population*. J. Am. Stat. Assoc., **80** (1985), 631–641.

[31] N. Wang, W.X. Zhang and C.L. Mei, *Fuzzy nonparametric regression based on local linear smoothing technique*. Information Sciences, **177** (2007), 3882–3900.

[32] M. Yang and H. Liu, *Fuzzy least squares algorithms for interactive fuzzy linear regression models*. Fuzzy Sets and systems, **135** (2003), 305–316.