Applicable Analysis and Discrete Mathematics
available online at http://pefmath.etf.rs

Appl. Anal. Discrete Math. **13** (2019), 583–604.
https://doi.org/10.2298/AADM171227021R

SPARSE REGULARIZED FUZZY REGRESSION

*Danilo Rapaić, Lidija Krstanović *, Nebojša Ralević, Ratko Obradović and Djuro Klipa*

In this work, we focus on two things: First, in addition to the data measurement uncertainty, we develop a novel probabilistic model by imposing the additive noise in the classical fuzzy regression model. We obtain the baseline LS estimation as the maximum likelihood estimation for regression parameters. Moreover, by assuming the heavy tail distribution and by introducing the Huber norm instead of square in the cost function, we obtain more general robust fuzzy M-estimator, much more suitable for modeling the outliers often present in the data sets.

1. Introduction

Classical regression models are used as a statistical tool in order to deliver the relationship between the independent and dependent variables, assuming that the difference between the observed and the dependent variables, which are crisp numbers, is due to the additive noise ε. When we assume the functional relationship to be affine, and additive noise to be Gaussian $\varepsilon \sim \mathcal{N}(0, \sigma)$, with the known standard deviation σ, the maximum likelihood estimate of crisp regression coefficients $b \in \mathbb{R}^d, d \in \mathbb{N}$ becomes an LS estimate. In the case of robust noise, as for example Laplace or heavy tailed additive noise, or the mixture of Gaussian and Heavy tailed additive noise, the robust M-estimators are introduced [**24, 25, 26**].

Fuzzy regression models are developed to construct the relationship between explanatory variables and response in a fuzzy environment. Those could be roughly

*Corresponding author. Lidija Krstanović
2010 Mathematics Subject Classification. 62A86, 62J86, 62J05
Keywords and Phrases. Fuzzy regression, sparse regularization, Huber norm, robust statistics, MAP estimate.

divided into: 1) Linear programming methods, [**4**]-[**8**]; 2) least-squares methods [**9**]-[**14**]; 3) support vector machines methods [**15**]. Fuzzy regression models are still in focus of many researchers (see [**36**]-[**40**], [**39**]). In [**36**] and Luo in [**37**], they use regularizer (mostly l_2) in some form on fuzzy regression coefficients. One of the most significant model was proposed by Tanaka in [**3**]. It has drawbacks in the sense that the larger amount of observations leads to fuzzier parameter estimates and makes the spread of the estimated fuzzy response wider [**4**], [**5**], which contradict the statistical principle that the larger amount of observed data produce lower variance of estimate, i.e., better estimate [**9**]. In order to deal with the mentioned problem, Diamond in [**11**] as well as Kao and Chyu in [**9**] adopted crisp instead of fuzzy regression coefficients. Many fuzzy regression models were proposed (see [**5**], [**10**], [**16**]) that use the criterion of Kim and Bishu, but the particular criterion has the property that the estimation error remains constant whenever the observed and the estimated response do not intersect with each other. In [**9**], Kao and Chyu also presented a mathematical programming model which minimizes the fuzzy error term corresponding to the sum of square errors between the observed and the estimated fuzzy response. That work, as well as those presented in for example [**14**], [**16**], [**11**] are nevertheless limited to fuzzy observations with triangular fuzzy numbers and are not computationally efficient. In [**17**] Chen and Hsueh, proposed a fuzzy regression approach that uses the LS method to minimize the total estimation error of the distance between the observed and estimated fuzzy responses. They presented each fuzzy observation as a fuzzy number and then they applied the LS method to determine the numeric regression coefficients and fuzzy adjustment variables to minimize the estimation error obtained on the basis of α-cuts of explanatory and observed variables. On the other hand, LASSO technique is well known in the field of classical, i.e., crisp regression, and is used in problems of prediction, as well as in feature selection [**18**] or recently [**47**] and [**48**]. It is obtained by imposing the l_1 sparse regularizer on regression coefficients into the particular cost that is to be minimized (see also [**41**], [**42**]). We mention that the sparsity property of the regression coefficients corresponds to the regression model where a small number of variables are actually really significant to the process of interest. In those cases, better estimates of the actual regression coefficients are obtained than in the case of classical LS estimates. The sparsity property in the problems of regression could be applied on a great number of real world problems [**18, 19**], also [**29, 30**] and most recently [**43**]-[**46**].

In this work we focus on two things. First, we observe that all the existing fuzzy regression models consider only the data measurement uncertainty, which are then modeled by using fuzzy numbers, but ignore the statistical nature of the noise which could be imposed on the channel that simulates the transmission path "between" the fuzzy measurements and the fuzzy response. As the result, the error term which figures in the cost function of most of the existing fuzzy regression models is based on the square error criterion, which is there imposed a-priori, basically mirrored from the crisp regression case, without appropriate probabilistic considerations. In line with that fact, we invoke the actual model of noise in the

channel that is "between" the fuzzy measurements and the fuzzy output, where we focus on the triangular fuzzy number case, which is sufficiently general[1]. We add the noise of the same fixed family of distributions on each of the parameters of the estimated fuzzy response. In that framework, the model proposed by Chen and Hsueh in [**17**][2] is the special case, when the actual noise imposed on the channel is Gaussian. Moreover, by using of the concepts of robust statistics, we further develop the proposed approach by using M-estimator and call it a Robust Fuzzy Regression (RFR) model. It deals with the problems of outliers, which are modeled by the usage of the Heavy tailed additive noise or the mixture of Heavy tailed and Gaussian additive noise. Our aim is to model the outliers in the robust way, but to still maintain existence and the uniqueness of the optimizer by keeping the convex nature of the optimization problem. Thus, we use Huber norm based M-estimator in the following sense: The robust error term of the cost that we propose is the sum of Huber norms of differences between the bounds of the α-cuts of the estimated and observed fuzzy variables, taken for all predefined α-cuts. Thus, we generalize the cost proposed in [**17**] by making it more robust to outliers. Second, we deliver the novel fuzzy regression model, in further text Sparse Regularized Fuzzy Regression (SRFR), which deals with the sparsity of the actual fuzzy data. It uses the assumption that most of the explanatory fuzzy variables have no significant influence on the actual response of the system. It is also delivered in the probabilistic framework as the Maximum a-posteriori (MAP) estimate of the regression coefficients, where we invoke the Laplace prior on the coefficients. The proposed model automatically distinguish significant from insignificant explanatory fuzzy variables in the model, without any prior knowledge on the nature of those particular variables. The cost function that we tend to minimize is contained of two parts: The first one is the error term between the observed and predicted fuzzy response, which uses the α-cuts of mentioned differences. The second one is the l_1 sparse regularizer on the crisp regression coefficients which tends to favor the sparseness of the data in comparison to the error term. Moreover, we further combine the Huber based M-estimator and the l_1 sparse regularizer and obtain Robust SRFR model in order to deal with the Heavy tailed noise imposed on the channel in the case of sparse fuzzy data.

The work is organized as follows: In Section 2 we give preliminaries, in Section 3, we introduce the probabilistic framework in fuzzy regression. In the same section we further introduce the novel RFR that deals with the heavy tailed noise imposed on the channel. In Section 4, we introduce the novel SRFR as well as Robust SRFR model, with the emphasis on l_1 regularization. Here we also explain how we use the method proposed by Wright *et al* [**20**] in order to obtain the solution to the proposed problem. In Section 5, we present experimental results on the synthetic fuzzy data.

[1] other types could also be delivered in the similar manner

[2] if we assume the triangular fuzzy numbers and some additional not to restrictive assumptions

2. Preliminaries

In this section we briefly elaborate on some basic definitions concerning fuzzy numbers with an accent on arithmetic operations on fuzzy numbers, which we later use in the paper.

A fuzzy number M is an normalized fuzzy set on $\mathbb{R}$, such that $\mu(x_0) = 1$ for only one $x_0 \in \mathbb{R}$, where piecewise continuous $\mu_M : \mathbb{R} \to \mathbb{R}$ is membership function of M. It is positive, iff $\mu_M|_{(-\infty,0]} \equiv 0$.

Recall, that for a fuzzy set A on $\mathbb{R}$, with membership function μ_A, α-cut, for $\alpha \in [0,1]$ is defined as $A_\alpha = \{x \in \mathbb{R} | \mu_A(x) \geq 0\}$. If A is fuzzy number, each α-cut is closed interval $[(A)^L_\alpha, (A)^R_\alpha]$ (see [**23**]).

Any of the four basic arithmetic operations on fuzzy numbers $* \in \{+, -, \cdot, /\}$ can be defined on fuzzy numbers A and B by using α-cut, by first defining

$$(A * B)_\alpha = A_\alpha * B_\alpha, \tag{1}$$

where $\alpha \in [0,1]$ and where $*$ on the right hand side of (1) is corresponding operation on closed intervals. For $[a,b], [d,e] \subset \mathbb{R}$, those are defined as

$$\begin{aligned}
&[a,b] + [d,e] = [a+d, b+e],\\
&[a,b] - [d,e] = [a-d, b-e],\\
&[a,b] \cdot [d,e] = [\min\{ad, ae, bd, be\}, \max\{ad, ae, bd, be\}]\\
&[a,b]/[d,e] = [\min\{a/d, a/e, b/d, b/e\}, \max\{a/d, a/e, b/d, b/e\}], \text{ provided } 0 \notin [d,e].
\end{aligned}$$

If M and N are fuzzy numbers, by application of the extension principle (see [**49**]) for the binary operation $* : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the membership function of the fuzzy number $M * N$ is given by

$$\mu_{M*N}(z) = \sup_{z=x*y} \min\{\mu_M(x), \mu_N(y)\} \tag{2}$$

A special type of representation for fuzzy numbers that we use in this paper is of the LR type (see [**49**]), as follows: Let $L(R) : \mathbb{R} \to [0,1]$ is a decreasing, with $L(0) = 1$, $L(x) < 1$, for $x > 0$, $L(x) > 0$, for $x < 1$ and $L(1) = 0$. Recall the following (see [**49**]):

Definition 1. *A fuzzy number M of LR-type, if there exist reference function L (for left), R (for right) and scalars $l, r > 0$ (left and right spreads), $m \in \mathbb{R}$ (mean) with*

$$\mu_M(x) = \begin{cases} L\left(\frac{m-x}{l}\right), & x < m \\ R\left(\frac{x-m}{r}\right), & x \geq m \end{cases} \tag{3}$$

It is denoted by $M = (m, l, r)_{LR}$.

Note, that by choosing $L(x) = max\{0, 1 - x\}$, we obtain triangular fuzzy number which are those that we further use in paper. Recall (see [**49**]), that for LR-type[3]fuzzy numbers $M = (m, l, r)_{LR}$, $N = (n, p, q)_{LR}$, the following definitions given in the LR representation framework is equivalent to corresponding definitions given by (2) ($* \in \{+, -\}$) and also to the definition (1):

$$\begin{aligned} M + N &= (m, l, r)_{LR} + (n, p, q)_{LR} = (m + n, l + p, r + q)_{LR} \\ M - N &= (m, l, r)_{LR} - (n, p, q)_{LR} = (m - n, l + q, r + p)_{LR}. \end{aligned} \tag{4}$$

Also, we obtain

$$-M = (0, 0, 0)_{LR} - M = (-m, r, l)_{LR} \tag{5}$$

and

$$\lambda M = \lambda(m, l, r)_{LR} = (\lambda m, |\lambda| l, |\lambda| r)_{LR},\ \lambda \in \mathbb{R}_+ \tag{6}$$

Triangular fuzzy numbers could also be presented by the triple $(u, m, v)_{TR}$, $u, m, v \in \mathbb{R}$, $u < m < v$, and by setting $l = m - u$, $r = v - m$, we obtain the equivalent LR representation $(m, l, r)_{LR}$. We say that the triangular fuzzy number in LR representation $(m, l, r)_{LR}$ is non-negative, iff it holds $m - l > 0$.

3. Probabilistic framework for fuzzy regression and the Robust Fuzzy Regression

One of the most effective methods to formulate the functional relationship between several input variables and the response variable is classical statistical regression, which is thus widely used in the feature selection problems. The model is formulated in the probabilistic, i.e., statistical manner (see for example [**21**]) in the following way. Let $(\mathcal{X}, \mathcal{Y}) = \{(X_{i1}, \ldots, X_{ip}; Y_i) | i = 1, \ldots, n\}$, $n \in \mathbb{N}$ be the set of observations, where X_{ij} represent the input variable of the ith observation, Y_i represents the response of the ith observation and n is the number of the observation. The model is given as:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_p X_{ip} + \varepsilon,\ i = 1, \ldots, n,\ n \in \mathbb{N}. \tag{7}$$

where $\boldsymbol{b} = [b_0, b_1, \ldots, b_p]^T \in \mathbb{R}^{p+1}$ are crisp regression coefficients and ε is zero mean additive Gaussian noise, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma)$, where σ is fixed known standard deviation. The term $\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_p X_{ip}$ is estimated while Y_i is

[3]Note that LR is just a representation of a fuzzy number.

observed response. It is well known (see for example [**21**]) that the ML estimate $\hat{\boldsymbol{b}}_{LS}$ is given by the LS estimate, which is obtained as the solution of the minimization problem $min \sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$.

Suppose that the observations, i.e., explanatory and response observed variables, are given as triangular fuzzy numbers $\{(\tilde{X}_{i1}, \ldots, \tilde{X_{ip}}; \tilde{Y}_i) | i = 1, \ldots, n\}$. We assume that by using those we can model the particular process with sufficient accuracy. The common fuzzy regression model used for example by [**3**]-[**6**] and [**17**], is then given by

$$\hat{\tilde{Y}}_i = \boldsymbol{b}^T \cdot \tilde{X}_i = b_0 + b_1\tilde{X}_{i1} + b_2\tilde{X}_{i2} + \cdots + b_p\tilde{X}_{ip},\ i = 1, \ldots, n,\ n \in \mathbb{N}. \tag{8}$$

where $\hat{\tilde{Y}}_i$ estimated fuzzy response and $\tilde{X}_i = [1\ \tilde{X}_{i1} \ldots \tilde{X_{ip}}]^T$. We note [**23**] that any linear combination (in means of operations on fuzzy numbers) of triangular fuzzy numbers, is a triangular fuzzy number. Term $b \in \mathbb{R}^p$, as in the case of classical regression model introduced preciously, represent crisp regression coefficients. In literature [**17**], [**22**], researchers used additional additive fuzzy adjustment term $\tilde{\delta}$ in order to deal with degenerate cases when the explanatory variables are crisp, so that the estimated response is also crisp, which leads to large fuzzy error for a given fuzzy response, but for the sake of simplicity and without loss of generality, we excluded that particular case, as our experimental data contain only "real" fuzzy explanatory observations, i.e., with no crisp observations. In order to invoke the statistical additive noise in the channel in conjunction with the fuzzy model (8), we formulate the following model

$$\tilde{Y}_i^t =_t \hat{\tilde{Y}}_i \oplus \tilde{\varepsilon},\ i = 1, \ldots, n,\ n \in \mathbb{N}. \tag{9}$$

where $\hat{\tilde{Y}}_i = \boldsymbol{b}^T \cdot \tilde{X}_i$ is estimated and $\hat{\tilde{Y}}_i$ is observed fuzzy response. In (9), term $=_t$ means that terms $\tilde{Y}_i^t$ on the left hand side of $=_t$ are only those realizations of random triple $\hat{\tilde{Y}}_i \oplus \tilde{\varepsilon}$ which are triangular fuzzy numbers, i.e., it holds $\tilde{Y}_i^t = (\tilde{Y}_l^{i,t}, \tilde{Y}_c^{i,t}, \tilde{Y}_r^{i,t})_{TR}$, with $\tilde{Y}_l^{i,t} < \tilde{Y}_c^{i,t} < \tilde{Y}_r^{i,t}$. In (9) the meaning of operation $\oplus$ is given by the Definition 2. We define operation $(\cdot) \oplus \tilde{\varepsilon}$ in the following way:

Definition 2. *Let $\varepsilon = (\varepsilon_l, \varepsilon_c, \varepsilon_r)$ be the noise triple where ε_h for $h \in \{l, c, r\}$ are independent random variables* [4] *representing additive noise corresponding to the left, center and right of particular triangular fuzzy number operand. Let $\tilde{X} = (X_l, X_c, X_r)_{TR}$, $X_l < X_c < X_r$ be a triangular fuzzy number. We define random triple $\tilde{Y} = (Y_l, Y_c, Y_r) = \tilde{X} \oplus \tilde{\varepsilon} = (X_l + \varepsilon_l, X_c + \varepsilon_c, X_r + \varepsilon_r)$.*

Remark 3. *Note that if ε_h are independent random variables, then $P((Y_l > Y_c) \cup (Y_c > Y_r)) > 0$, so there is no guaranty that the realizations of $\tilde{Y}_i$ are fuzzy numbers. The sort operation performed on realizations of $\tilde{Y}_h^i$ $h \in \{l, c, r\}$ that could be invoked*

[4] Two random variables X and Y are independent if their probability distributions satisfy $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

in order to assure that the realizations of $\tilde{Y}_i$ are indeed triangular fuzzy numbers, would spoil the Gaussian (or Cauchy) distribution of Y_i which is the assumption (see Propositions 1 and 2) that we use further assume. In the case when the noise triple satisfies $\varepsilon_h = \varepsilon$, $h = l, c, r$, i.e., all three components are the same random variable, then every realization of the triple $\tilde{Y} = (Y_l, Y_c, Y_r) = \tilde{X} \oplus \varepsilon$ is indeed triangular fuzzy number.

We deliver the following observations which correspond to models that use assumptions on Gaussian and Cauchy additive noise models respectively, as well as the mixture of the previous two, which we use further in formulating our target cost functionals.

Proposition 1. *Let $\tilde{X} = (X_l, X_c, X_r)$ be a triangular fuzzy number. Let $\tilde{Y} = \tilde{X} \oplus \tilde{\varepsilon}$, where $(\cdot) \oplus \tilde{\varepsilon}$ is defined by definition 2 and $\varepsilon = (\varepsilon_l, \varepsilon_c, \varepsilon_r)$, where $\varepsilon_h \sim \mathcal{N}(0, \sigma_h)$, for $h \in \{l, c, r\}$ are zero mean Gaussian variables. Let $(\tilde{Y})^L_\alpha = (1-\alpha)Y_l + \alpha Y_c$, $(\tilde{Y})^R_\alpha = (1-\alpha)Y_c + \alpha Y_r$. Then it holds $(\tilde{Y})^L_\alpha \sim \mathcal{N}((1-\alpha)X_l + \alpha X_c, \sqrt{(1-\alpha)^2\sigma_l^2 + \alpha^2\sigma_c^2})$ and also $(\tilde{Y})^R_\alpha \sim \mathcal{N}((1-\alpha)X_c + \alpha X_r, \sqrt{(1-\alpha)^2\sigma_c^2 + \alpha^2\sigma_r^2})$.*

Proof:

We consider the case $(\tilde{Y})^L_\alpha$ and the proof for case $(\tilde{Y})^R_\alpha$ is analogous. It holds that $\tilde{Y}_l = X_l + \varepsilon_l \sim \mathcal{N}(X_l, \sigma_l)$ and $\tilde{Y}_c = X_c + \varepsilon_c \sim \mathcal{N}(X_c, \sigma_c)$, so that any of their convex combination also have Gaussian distribution. For convex combination $Z_\lambda = \lambda X_l + (1-\lambda)X_c + \lambda\varepsilon_l + (1-\lambda)\varepsilon_c$ for some $\lambda \in [0,1]$, we obtain $E(Z_\lambda) = \lambda X_l + (1-\lambda)X_c$ and $D(Z_\lambda) = D(\lambda X_l + (1-\lambda)X_c) + D(\lambda\varepsilon_l + (1-\lambda)\varepsilon_c) = \lambda^2\sigma_l^2 + (1-\lambda)^2\sigma_c^2$. As it holds $(\tilde{Y})^L_\alpha = (1-\alpha)\tilde{Y}_l + \alpha\tilde{Y}_c$, based on the previous we obtain that $(\tilde{Y})^L_\alpha$ is drown from $\mathcal{N}((1-\alpha)X_l + \alpha X_c, \sqrt{(1-\alpha)^2\sigma_l^2 + \alpha^2\sigma_c^2})$, it ends the proof. □

Proposition 2. *Let $\tilde{X} = (X_l, X_c, X_r)$ be a triangular fuzzy number. Let $\tilde{Y} = \tilde{X} \oplus \tilde{\varepsilon}$, where $(\cdot) \oplus \tilde{\varepsilon}$ is defined by definition 2 and $\varepsilon = (\varepsilon_l, \varepsilon_c, \varepsilon_r)$, where $\varepsilon_h = \mathcal{C}(0, \gamma_h) = \frac{1}{\pi}\left[\frac{1}{1+(x/\gamma_h)^2}\right]$, for $h \in \{l, c, r\}$ are zero median Cauchy variables. Let $(\tilde{Y})^L_\alpha = (1-\alpha)Y_l + \alpha Y_c$ and $(\tilde{Y})^L_\alpha = (1-\alpha)Y_c + \alpha Y_r$. Then it holds $(\tilde{Y})^L_\alpha \sim \mathcal{C}((1-\alpha)X_l + \alpha X_c, (1-\alpha)\gamma_l + \alpha\gamma_c)$ and also $(\tilde{Y})^R_\alpha \sim \mathcal{C}((1-\alpha)X_c + \alpha X_r, (1-\alpha)\gamma_c + \alpha\gamma_r)$.*

Proof:

We consider the case $(\tilde{Y})^L_\alpha$ and the proof for case $(\tilde{Y})^R_\alpha$ is analogous. We consider the convex combination

$$
\begin{aligned}
(10) \qquad Z_\lambda = \lambda Y_l + (1-\lambda)Y_c &= \lambda(X_l + \varepsilon_l) + (1-\lambda)(X_c + \varepsilon_c) \\
&= (\lambda X_l + (1-\lambda)X_c) + \lambda\varepsilon_l + (1-\lambda)\varepsilon_c \\
&= A + Z^\varepsilon_\lambda
\end{aligned}
$$

for any fixed $\lambda \in [0,1]$, where $A = \lambda X_l + (1-\lambda)X_c$ and $Z^\varepsilon_\lambda = \lambda\varepsilon_l + (1-\lambda)\varepsilon_c$. It holds that if $X \sim \mathcal{C}(0, \gamma_l)$, then $X + x_0 \sim \mathcal{C}(x_0, \gamma_l)$ (it could be easily proved by using characteristic functions). Using the fact that ε_l and ε_c are independent, which

implies $p(\varepsilon_l\varepsilon_c) = p(\varepsilon_l)p(\varepsilon_c)$ and also the fact that $\varepsilon_l \sim \mathcal{C}(0,\gamma_l)$ and $\varepsilon_c \sim \mathcal{C}(0,\gamma_c)$ respectively, the characteristic function $k_{Z_\lambda}(t)$ of Z_λ is equal to

$$
\begin{aligned}
(11) \qquad k_{Z^\varepsilon_\lambda}(t) &= E[e^{i\lambda\varepsilon_l t}e^{i(1-\lambda)\varepsilon_c t}] \\
&= \int_{-\infty}^{+\infty} e^{i\lambda\varepsilon_l t}p(\varepsilon_l)d\varepsilon_l \int_{-\infty}^{+\infty} e^{i(1-\lambda)\varepsilon_c t}p(\varepsilon_c)d\varepsilon_c \\
&= E[e^{i\lambda\varepsilon_l t}]E[e^{i(1-\lambda)\varepsilon_c t}] = k_{\lambda\varepsilon_l}(t)k_{(1-\lambda)\varepsilon_c}(t)
\end{aligned}
$$

As it holds that $k_{\lambda\varepsilon_l}(t) = e^{-\lambda\gamma_l|t|}$ and $k_{(1-\lambda)\varepsilon_c}(t) = e^{-(1-\lambda)\gamma_c|t|}$ respectively (it can be easily verified by using characteristic functions), from (11) we obtain $k_{Z^\varepsilon_\lambda}(t) = e^{-(\lambda\gamma_l+(1-\lambda)\gamma_c)|t|}$, and consequently $Z^\varepsilon_\lambda \sim \mathcal{C}(0,\lambda\gamma_l + (1-\lambda)\gamma_c)$. Now, as $Z_\lambda = A + Z^\varepsilon_\lambda$, it holds that $Z_\lambda \sim \mathcal{C}(A,\lambda\gamma_l + (1-\lambda)\gamma_c)$. As it holds $(\tilde{Y})^L_\alpha = (1-\alpha)\tilde{Y}_l + \alpha\tilde{Y}_c$, from previous, we obtain that $(\tilde{Y})^L_\alpha$ is drown from $\mathcal{C}((1-\alpha)X_l + \alpha X_c, (1-\alpha)\gamma_l + \alpha\gamma_c)$, which ends the proof.

□

Using the similar arguments as in Propositions 1 and 2, it can be shown that the similar can be concluded for mixtures of Gaussian and Cauchy noise:

Corollary 1. *Let $\tilde{X} = (X_l, X_c, X_r)$ be a triangular fuzzy number. Let $\tilde{Y} = \tilde{X} \oplus \tilde{\varepsilon}$, where $(\cdot) \oplus \tilde{\varepsilon}$ is defined by definition (2) and $\varepsilon = (\varepsilon_l, \varepsilon_c, \varepsilon_r)$, where $\varepsilon_h = \alpha_h\varepsilon^g_h + (1-\alpha_h)\varepsilon^{ch}_h$, with $\varepsilon^g_h \sim \mathcal{N}(0,\sigma_h)$, $\varepsilon^{ch}_h \sim \mathcal{C}(0,\gamma_h)$, for $h \in \{l,c,r\}$, $\alpha_h \in [0,1]$. Let $(\tilde{Y})^L_\alpha = (1-\alpha)Y_l + \alpha Y_c$ and $(\tilde{Y})^L_\alpha = (1-\alpha)Y_c + \alpha Y_r$. Then $(\tilde{Y}^t)^L_\alpha = Z^g_1 + Z^{ch}_1$, where $Z^g_1 \sim \mathcal{N}((1-\alpha)X_l + \alpha X_c, \sqrt{(1-\alpha)^2\alpha^2_l\sigma^2_l + \alpha^2\alpha^2_c\sigma^2_c})$ and $Z^{ch}_1 \sim \mathcal{C}(0,(1-\alpha)(1-\alpha_l)\gamma_l + \alpha(1-\alpha_c)\gamma_c)$. Also, it holds $(\tilde{Y})^R_\alpha = Z^g_2 + Z^{ch}_2$, where Z^g_2 is drown from $\mathcal{N}((1-\alpha)X_c + \alpha X_r, \sqrt{(1-\alpha)^2\alpha^2_c\sigma^2_c + \alpha^2\alpha^2_r\sigma^2_r})$ and Z^{ch} is drown from $\mathcal{C}(0,(1-\alpha)(1-\alpha_c)\gamma_c + \alpha(1-\alpha_r)\gamma_r)$.*

Remark 4. *As it is already noted in Remark 3, the fuzzy data $\tilde{Y}^t$ that we deal with will (see (9) and the explanation above) be composed of triangular fuzzy numbers which we consider to be realizations of $\tilde{Y}$ present in Propositions 1, 2 and Corollary 1, such that it holds $\tilde{Y}^t = (\tilde{Y}^t_l, \tilde{Y}^t_c, \tilde{Y}^t_r)$ with $\tilde{Y}^t_l < \tilde{Y}^t_c < \tilde{Y}^t_r$ satisfied, i.e., it holds that $(\tilde{Y}^t_l, \tilde{Y}^t_c, \tilde{Y}^t_r)_{TR}$ is triangular fuzzy number. Thus, by doing that, we actually distort the particular distribution listed in 1, 2 and Corollary 1, so that in the rest of the paper, for an arbitrary α-cut $[(\tilde{Y}^t_i)^L_\alpha, (\tilde{Y}^t_i)^R_\alpha]$ of $\tilde{Y}^t_i$, $i = 1,\ldots,n$, for some $\alpha \in [0,1]$, we assume that $(\tilde{Y}^t)^L_\alpha$ and $(\tilde{Y}^t)^L_\alpha$ are approximately distributed as in some of the previously mentioned propositions.*

We further give the statistical interpretation of cost function invoked in work of Chen and Hsueh [**17**], which is a staring point for our robust fuzzy model as well as our sparse representation and the fusion of the two.

Let us consider a given set of fuzzy observations $(\mathcal{X}, \mathcal{Y}^t)$ given as follows: Let $\mathcal{X} = \{\tilde{X}_i | i = 1,\ldots,n\}$ with $\tilde{X}_i = (\tilde{X}_{ij} | j = 1,\ldots,p)$ where $\tilde{X}_{ij}$ are triangular fuzzy numbers. Let $\mathcal{Y}^t = \{\tilde{Y}^t_i | i = 1,\ldots,n\}$, where $\tilde{Y}^t_i = (Y^{(i),t}_l, Y^{(i),t}_c, Y^{(i),t}_r)_{TR} =_t \tilde{Y}_i$ are

realizations of i.i.d. random triples $\tilde{Y}_i = \hat{\tilde{Y}}_i \oplus \varepsilon$, $\hat{\tilde{Y}}_i = \boldsymbol{b}^T \cdot \tilde{X}_i$, with $\oplus$ given by (9) and Definition 2, with the additional assumption that those realizations are triangular fuzzy numbers, i.e., $Y_l^{(i),t} \leq Y_c^{(i),t} \leq Y_r^{(i),t}$ holds (see Remark 3). Consider also a given set $\Lambda_k = \{\alpha_1, \ldots, \alpha_k\}$, $\alpha_j \in [0,1]$, $j = 1, \ldots, k$.

Because of the previously mentioned independence assumption, it holds for the data likelihood [5]

$$
\begin{aligned}
& p(\mathcal{X}, \mathcal{Y}^t | \mathbf{b}) = p((\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_1,L}, (\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_1,R}, \ldots, (\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_k,L}, (\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_k,R}, \\
& \ldots, \\
& (\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_1,L}, (\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_1,R}, \ldots, (\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_k,L}, (\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_k,R}) = \\
& p((\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_1,L}, \ldots, (\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_k,L}) \\
& p((\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_1,R}, \ldots, (\tilde{Y}_1^t - \hat{\tilde{Y}}_1)^{\alpha_k,R}) \\
& p((\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_1,L}, \ldots, (\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_k,L}) \\
& p((\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_1,R} \ldots, (\tilde{Y}_N^t - \hat{\tilde{Y}}_N)^{\alpha_k,R})
\end{aligned}
\tag{12}
$$

and further, as for any sequence of random variables $\{X_i\}_1^n$ it holds $p(X_n, \ldots, X_1) = p(X_n | X_{n-1} \ldots X_1) \cdots p(X_2|X_1)p(X_1)$, we have

$$
\begin{aligned}
& p((\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_1,L}, \ldots, (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_k,L}) \\
& = p((\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_1,L} | (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_2,L}, \ldots, (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_k,L}) \\
& \quad \ldots \\
& p((\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_2,L} | (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_1,L}) p((\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_1,L})
\end{aligned}
\tag{13}
$$

If we assume $\varepsilon_h \sim \mathcal{N}(0, \sigma_h)$, for $h \in \{l, c, r\}$, then based on Proposition 1 and Remark 4, for conditional random variables appearing in (13) it holds:

$$
(\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_q,L} | (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_{q+1},L}, \ldots, (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_k,L} \approx \mathcal{N}\left(0, \sqrt{(1-\alpha_q)^2\sigma_l^2 + \alpha_q^2\sigma_c^2}\right), \tag{14}
$$

as well

$$
(\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_q,R} | (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_{q+1},R}, \ldots, (\tilde{Y}_i^t - \hat{\tilde{Y}}_i)^{\alpha_k,R} \approx \mathcal{N}\left(0, \sqrt{(1-\alpha_q)^2\sigma_c^2 + \alpha_q^2\sigma_r^2}\right), \tag{15}
$$

where $q = 1, \ldots, k$ and use symbol $\approx$ is instead of $\sim$ to emphasize the fact that this is approximate, because we neglect situations when $\tilde{Y}_l^{(i),t} \leq \tilde{Y}_c^{(i),t} \leq \tilde{Y}_r^{(i),t}$ are not satisfied.

[5] Note that in (12), $\mathbf{b}$ is present implicitly via $\hat{\tilde{Y}}_i$

In the special case $\varepsilon_h \sim \mathcal{N}(0,\sigma),\ h=l,c,r$ with fixed standard deviation σ, by taking the maximum likelihood approach, i.e., by applying the $\ln(\cdot)$ operation of likelihood (12) and by taking into account (12)-(15), we obtain that the following cost

$$\mathcal{L}(\boldsymbol{b}) = \sum_{i=1}^{n} \sum_{\alpha\in\Lambda_k} \left[((\hat{\tilde{Y}}_i^L)_\alpha - (\tilde{Y}^L)_\alpha^t)^2 + ((\hat{\tilde{Y}}_i^R)_\alpha - (\tilde{Y}^R)_\alpha^t)^2 \right] \tag{16}$$

which is to be minimized with respect to the regression parameters $\boldsymbol{b} \in \mathbb{R}^{p+1}$ in order to obtain the ML estimate. The cost (16) is actually proposed and used by Chen and Hsueh in [**17**]. The solution $\boldsymbol{b}_{LS}$ that minimizes (16) is delivered in the close form in [**17**].

Nevertheless, in the presence of outliers, i.e., atypical observations that differ from the main part of the data set, the Heavy tailed additive noise is usually used instead of Gaussian in the actual regression model (see [**24, 25**]). It is for example, common approach in image processing, and can be induced during the acquisition process [**31, 32**]. It is also common to model outliers with the mixture of Heavy tailed and Gaussian noise (see for example applications in image processing [**31, 33**]). In order to deal with problems when there is an assumption on Heavy tailed additive noise, an M-estimators are introduced in the area of ordinary i.e. crisp linear regression, which utilize robust statistics approach [**24**]. To the knowledge of authors of this paper, such an approach is not applied yet in the area of fuzzy regression modeling. The main reason is, that in the existing literature, there is no proper statistical interpretation of costs used in fuzzy regression in the means of the noise imposed on the channel which is one of the novelty that this work introduces. We go further by introducing the cost that can be viewed as fuzzy M-estimator that deals with the situation where the additive noise $\tilde{\varepsilon}$ imposed on the channel (see the definition (2)) is such that $\varepsilon_h,\ h=l,c,r$ is robust noise of Cauchy type, or the mixture of Cauchy and Gaussian type, which are one of the most common of robust noise models used in applications. Thus, based on Proposition 2 and Corollary 1, we conclude that conditional variables appearing at the left-hand side of (14) and (15) both have either Cauchy, or the mixture of Cauchy and the Gaussian distributions (with corresponding parameters), respectively, instead of Gaussian distribution as it in (14) and (15). Thus, as the fuzzy M-estimator we propose the one that utilize the Huber norm on each α-cut. It is most commonly used as M-estimator for that particular types of robust noise, which is due to its robustness and simplicity (see [**24**]). In the line with previous, we formulate the following cost

$$\mathcal{L}(\boldsymbol{b}) = \sum_{i=1}^{n} \sum_{\alpha\in\Lambda_k} \rho_m((\hat{\tilde{Y}}^L)_\alpha - (\tilde{Y}^L)_\alpha^t) + \rho_m((\hat{\tilde{Y}}^R)_\alpha - (\tilde{Y}^R)_\alpha^t) \tag{17}$$

which is to be minimized with respect to $\boldsymbol{b} \in \mathbb{R}^{p+1}$. The term $\rho_m(\cdot),\ m>0$, is the Huber norm defined by

$$\rho_m(x) = \begin{cases} \frac{x^2}{2m}, & 0 \leq |x| \leq m \\ |x| - \frac{m}{2}, & |x| > m \end{cases} \tag{18}$$

We call the proposed method that utilized the cost (17), the Robust Fuzzy Regression (RFR). As (18) is a convex function with respect to the $\boldsymbol{b}$, it stays convex as in the case of (17), so that consequently, the existence and the uniqueness of the minimizer is maintained. As $\rho_m \in C^1(\mathbb{R})$ for all $m > 0$, it is obvious that, for example, simple gradient descent methodology could be applied in order to minimize cost (17). The choice of parameter m must be such that it is sufficiently large in order not to mask the Gaussian like behavior of the data near the origin, and at the same time sufficiently small in order to catch outliers. We obtain it heuristically, dependent on the actual application. In our experiments, as we target sparse regularized case that we propose, we use the method proposed by Wright et al. in [**20**] to efficiently obtain optimal solution to the corresponding sparse regularized minimizing problems. For the simplicity we use the same method in order to obtain the minimizer for the cost (17) by just setting $\tau = 0$ (see [**20**]). We note that M-estimators, such as Huber estimator which we use in this paper, are able to interpolate between l_1 and l_2 estimators (by using different values of $\varepsilon > 0$) and thus potentially obtain better results on some particular data set. Moreover, the convex combination $\lambda\|\cdot\|_{l_1} + (1-\lambda)\|\cdot\|_{l_2}$, $\lambda \in [0,1]$ could also interpolate between l_1 and l_2 estimators, but we did not considered those possibilities in this work.

4. Sparse Regularized Fuzzy Regression

One of the major problems that emerges in the application of the regression models, crisp or fuzzy, on the real world problems is that the set of the input or explanatory variables is chosen with out enough prior knowledge of the actual problem that one tends to model. In many cases, it results in the fact that there are to many input variables and that most of them have no significant influence on the output variable, i.e., the response of the system, thus worsening the performance of the regression. Equivalently, it means that only few of them have significant influence on the response. It is referred to, as the sparsity of the data. In the area of crisp regression the problem of feature reduction which extract those significant variables is specially effectively solved by using the LASSO regression models, first introduced by Tibshirani in [**18**], and further developed in [**34, 35**].

Although, the same previously described problem also emerges in the area of fuzzy regression and was treated by using l_2 regularizer (see [**36**] and [**37**]), to the knowledge of authors of this paper, it was not treated by the means of sparse regularizer of any kind. In line with that, in this section we expand the concept of LASSO which is well known in the crisp regression, on the problems of fuzzy regression. We do it by using the probabilistic framework that we impose in the previous

section which enables us to consider classical LS as well as robust M-estimator in conjunction with the actual l_1 regularizer which emerge as a consequence of the Laplace prior imposed on the crisp regression coefficients $\boldsymbol{b}$. Actually, we go along the line of Bayesian approach and assume that the vector of crisp regression parameters $\boldsymbol{b}$ is random vector with probability density $p(\boldsymbol{b})$. In order to obtain the MAP estimate of $\boldsymbol{b}$, we tend to maximize a-posterior probability density $p(\boldsymbol{b}|(\mathcal{X}, \mathcal{Y}^t))$. Using the Bayesian theorem, we obtain

$$\begin{aligned} \boldsymbol{b}_{MAP} &= \operatorname{argmax}_{\boldsymbol{b}}\, p((\mathcal{X}, \mathcal{Y}^t)|\boldsymbol{b})p(\boldsymbol{b}) \\ &= \operatorname{argmax}_{\boldsymbol{b}}\, \left[\ln(p((\mathcal{X}, \mathcal{Y}^t)|\boldsymbol{b})) + \ln p(\boldsymbol{b})\right] \end{aligned} \tag{19}$$

with $p((\mathcal{X}, \mathcal{Y}^t)|\boldsymbol{b})$ given by (12).

In order to obtain sparse regularization in the final cost, we assume the Laplace prior on coefficients $\boldsymbol{b}$, i.e., that $\boldsymbol{b} \sim p(\boldsymbol{b}) = (1/2\eta)e^{-\|\boldsymbol{b}\|_1/\eta}$, with fixed $\eta > 0$. The term $\|\boldsymbol{b}\|_1$ is l_1 norm of parameters $\boldsymbol{b} \in \mathbb{R}^{p+1}$, defined by $\|\boldsymbol{b}\|_1 = \sum_{k=0}^{p} |b_k|$. If we assume the Gaussian case in the sense of (2), i.e., $\varepsilon_h = \varepsilon,\ h = l, c, r$ is zero mean Gaussian noise given by $p_\varepsilon(y|\Theta) = \varepsilon \sim \mathcal{N}(0, \sigma)$ with fixed standard deviation σ, and simplifying $\Theta_{L,\alpha} \approx \sigma,\ \Theta_{R,\alpha} \approx \sigma$ and also setting $\lambda = 1/\eta$, we obtain the MAP estimate as the solution to the task of minimizing the following cost:

$$\mathcal{L}(\boldsymbol{b}) = \sum_{i=1}^{n} \sum_{\alpha \in \Lambda_k} \left[(\hat{(\tilde{Y}^L)}_\alpha - (\tilde{Y}^L)^t_\alpha)^2 + (\hat{(\tilde{Y}^R)}_\alpha - (\tilde{Y}^R)^t_\alpha)^2 \right] + \lambda \|\boldsymbol{b}\|_1 \tag{20}$$

with respect to regression parameters $\boldsymbol{b} \in \mathbb{R}^{p+1}$. The cost (20) is the generalization of the cost (16), by the means of l_1 sparse regularization, which is to be applied if the assumption on fuzzy data sparsity holds. We call the proposed method that utilize cost (20) the Sparse Fuzzy Regression (SFR).

In the robust noise case in the sense of (2), if we assume $\varepsilon_h,\ h = l, c, r$ is robust additive noise of Cauchy type, or the mixture of Cauchy and Gaussian type of noise, if the additional assumption on fuzzy data sparsity is imposed, we use the fuzzy M-estimator which we introduce in the previous section in conjunction with the l_1 sparse regularization corresponding to the Laplace prior with parameter η. Thus, we model the combine case of fuzzy data sparsity coupled with the presence of data outliers inherent to the robustness of the noise imposed on the data, where robust fuzzy M-estimator also obtain significantly better results in comparison to LS, as we show on our experiments in Section 5. We obtain the task of minimizing the following cost

$$\mathcal{L}(\boldsymbol{b}) = \sum_{i=1}^{n} \sum_{\alpha \in \Lambda_k} \rho_m(\hat{(\tilde{Y}^L)}_\alpha - (\tilde{Y}^L)^t_\alpha) + \rho_m(\hat{(\tilde{Y}^R)}_\alpha - (\tilde{Y}^R)^t_\alpha) + \lambda \|\boldsymbol{b}\|_1 \tag{21}$$

with respect to parameters $\boldsymbol{b} \in \mathbb{R}^{p+1}$, where ρ_m is Huber norm defined by (18). The value of parameter $m > 0$ is obtained heuristically, depending on the actual

application, as in the case of (17) in the previous section. Parameter $\lambda = 1/\eta$ is obtained as in the case of (20). We call the proposed method that utilize cost (21), the Robust Sparse Fuzzy Regression (RSFR).

The costs (20) and (21) are convex, continuous and coercive with respect to parameters $\boldsymbol{b} \in \mathbb{R}^{p+1}$, which implies that in both cases, from arbitrary minimizing sequence, one can extract subsequence converging to the unique minimizer (see [**28**]). Note however, that both (20) and (21) contain regularization term $\|\boldsymbol{b}\|_1$ which is not smooth, so that smooth optimization techniques can not be implied in order to obtain their minimizers. For that task we use the SpaRSA method proposed by Wright et al. in [**20**]. It is the method for solving general unconstrain optimization problem of the form $\min_x \phi(x) := f(x) + \lambda c(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is smooth function, and $c : \mathbb{R}^n \to \mathbb{R}$ is regulizer, usually non-smooth (as it is the case in our application) and possibly non-convex. The case of regularizer $\|\boldsymbol{b}\|_1 = \sum_{k=0}^{p} |b_k|$ which we use in the cost (20) and (21) is the special convex case of the more general separable regularizer $c(\boldsymbol{b}) = \sum_{k=0}^{p} c_k(b_k)$, where $c_k(\cdot)$ are arbitrary non-smooth and possibly non-convex real functions of the real argument. Following the SpaRSA method [**20**], in the case of cost (20), we set

$$f_1(\boldsymbol{b}) = \sum_{i=1}^{n} \sum_{\alpha \in \Lambda_k} \left[((\hat{\tilde{Y}}^L)_\alpha - (\tilde{Y}^L)^t_\alpha)^2 + ((\hat{\tilde{Y}}^R)_\alpha - (\tilde{Y}^R)^t_\alpha)^2 \right] \tag{22}$$

and in the case of (21), we set

$$f_2(\boldsymbol{b}) = \sum_{i=1}^{n} \sum_{\alpha \in \Lambda_k} \rho_m((\hat{\tilde{Y}}^L)_\alpha - (\tilde{Y}^L)^t_\alpha) + \rho_m((\hat{\tilde{Y}}^R)_\alpha - (\tilde{Y}^R)^t_\alpha). \tag{23}$$

for which we obtain gradients in closed form. In the case of cost (20), we obtain j-th component of $\nabla f_1(\boldsymbol{b})$ as

$$\begin{aligned} \frac{\partial}{\partial b_j} f_1(\boldsymbol{b}) &= \sum_{i=1}^{n} \sum_{\alpha \in \Lambda_k} 2((\hat{\tilde{Y}}^L)_\alpha - (\tilde{Y}^L)^t_\alpha)(X^L_{ij})_\alpha \\ &+ 2((\hat{\tilde{Y}}^R)_\alpha - (\tilde{Y}^R)^t_\alpha)(X^R_{ij})_\alpha \end{aligned} \tag{24}$$

and also of $\nabla f_2(\boldsymbol{b})$ as

$$\begin{aligned} \frac{\partial}{\partial b_j} f_2(\boldsymbol{b}) &= \sum_{i=1}^{n} \sum_{\alpha \in \Lambda_k} \psi_m((\hat{\tilde{Y}}^L)_\alpha - (\tilde{Y}^L)^t_\alpha)(X^L_{ij})_\alpha \\ &+ \psi_m((\hat{\tilde{Y}}^R)_\alpha - (\tilde{Y}^R)^t_\alpha)(X^R_{ij})_\alpha \end{aligned} \tag{25}$$

where ψ_m is a derivative of the Huber norm with parameter m, given by

$$\psi_m(x) = \begin{cases} \frac{x}{m}, & 0 \leq |x| \leq m \\ 1, & x > m \\ -1, & x < m \end{cases}, \tag{26}$$

and f_1 and f_2 are functions of $\boldsymbol{b}$ via the terms $\hat{\tilde{Y}}^L$ and $\hat{\tilde{Y}}^R$.

Now, following the line of SpaRSA [**20**], we obtain minimizers for cost (20) and cost (21) respectively by alternatingly performing following two steps: Step one is given by

$$\boldsymbol{b}^{t+1} = \underset{z}{\operatorname{argmin}} \frac{1}{2}(z - u_i^t)^2 + \frac{\lambda|z|}{\alpha_t} = \operatorname{soft}\left(u_i^t, \frac{\lambda}{\alpha_t}\right) \tag{27}$$

where $\operatorname{soft}(u, a) \equiv \operatorname{sign}(u) \max\{|u| - a, 0\}$ denotes the well known soft-threshold function. Step two is given by

$$\boldsymbol{u}^t = \boldsymbol{b}^t - \frac{1}{\alpha_t} \nabla f_1(\boldsymbol{b}^t) \tag{28}$$

in the case of cost (20), and by

$$\boldsymbol{u}^t = \boldsymbol{b}^t - \frac{1}{\alpha_t} \nabla f_2(\boldsymbol{b}^t) \tag{29}$$

in the case of cost (21) respectively. We obtain α_t adaptively, following the SpaRSA scheme proposed in [**20**]. Under the mild condition (see [**20**]) which are all satisfied for both cost (20) and (21), this scheme obtain stationary point which is in our case the unique solution of the corresponding minimization problems.

5. Experimental results

In this Section we give the experimental results on the syntectic data. We evaluate the performance of the proposed RFR, SFR and RSFR respectively. The experiments are given, in comparison to the baseline FR model given by Chen and Hsueh in [**17**] in the case of non-sparse and sparse fuzzy data, and in the presence of Gaussian, Cauchy noise and the mixture of these two. Moreover, we compare proposed regression models among each other, for the same data settings.

For all experiments, elements of the observation set $\mathcal{Y} = \{\tilde{Y}_i | i = 1 \ldots, n\}$ are generated as the random triangular fuzzy numbers $\tilde{Y}_i = \hat{\tilde{Y}}_i \oplus \tilde{\varepsilon}$ by the means of (9) and the definition (2), where $\varepsilon_h \sim p(\cdot, \Theta_h),\ h \in \{l, c, r\}$ is Gaussian, Cauchy or the

mixture of these two distributions, depending on the particular experiment. We obtained $\hat{\tilde{Y}}_i = \boldsymbol{b}^T \cdot \tilde{X}_i$, where for the particular experiment, we fix the dimension $p \in \mathbb{N}$ and the true regression parameters $\boldsymbol{b} \in \mathbb{R}^{p+1}$ were given a priori and the task of the experiments was to reconstruct them with the minimal possible error. The input variables $\tilde{X}_i$ also randomly generated for the sake of completeness of the experiments, although are considered in the previous sections to be given as deterministic input. For each $i \in \{1, \ldots, n\}$, the center of the triangular fuzzy number $\tilde{X}_i$ is generated by setting $X_{i,c} = LS + f$, where $S \sim \mathcal{U}(0, 1)$ and we set $L = 10,\ f = 1$. For the simplicity and without loosing generality, we obtain $X_{i,l} = X_{i,c} - D_l$, $X_{i,r} = X_{i,c} + D_r$, where $D_l = 0.5$, $D_r = 0.2$ are fixed. Number of α-cuts was $k = 3$, for all experiments. In each experiment we estimate $\hat{\boldsymbol{b}}$ and report the mean square error, calculated as $E_{sqr} = \frac{1}{p+1}\|\hat{\boldsymbol{b}} - \boldsymbol{b}\|_2^2$, for different parameter setting. In order to obtain minimizer for the baseline FR, we use the close form solution proposed in [**17**], in all experiments. In order to obtain minimizers in case of the proposed RFR, SFR and RSFR, we use the SpaRSA algorithm [**20**], where we set $\alpha_0 = 0.07$, $\alpha_{min} = 0.05$, $\alpha_{max} = 0.2$, $\eta = 1.3$ in all cases. In the cases of SFR and RSFR, we set $\tau = 20$, while in the RFR case we set $\tau = 0$, in all the experiments. Also, in the cases of RFR and RSFR, we set $m = 10$ for the parameter of Huber norm ρ_m, in all experiments. For each experiment, we vary the following parameters: number of observations n, scale parameter $\nu > 0$ of the Cauchy noise and the standard deviation $\sigma > 0$ of the Gaussian noise, if they are used in the particular experiment.

In Table 1, experimental results of comparison of the proposed RFR with the baseline FR model are presented. We used fixed $p = 7$, and also the true regression vector $\boldsymbol{b} = [0.4, 5, 8, 2, 3.6, 7.2, 0.2, 0.1]$, while varying the scale of the Cauchy noise $\nu > 0$ and the number of observations n. It can be seen that in the case of Gaussian noise, the proposed RFR obtain significantly better results in the means of mean square error E_{sqr}, in comparison to the baseline FR model, for all scales ν and all number of observations n tested.

In Table 2, experimental results of comparison of the proposed SFR with the baseline FR model are presented. We used fixed $p = 19$, and also the true regression vector $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0]$, $dim(\boldsymbol{b}) = 20$, while varying the standard deviation of the Gaussian noise σ and the number of observations n. It can be seen that in the case of sparse fuzzy data and Gaussian noise, the proposed SFR obtain significantly better results in the means of mean square error E_{sqr}, in comparison to the baseline FR model, for all standard deviations σ and all number of observations n tested.

In Table 3, experimental results of comparison of the proposed SFR and RSFR are presented. We used fixed $p = 19$, and also the true regression vector $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0]$, $dim(\boldsymbol{b}) = 20$, while varying the scale parameter ν of the Cauchy noise imposed, and the number of observations n. It can be seen that in the case of sparse fuzzy data and imposed Cauchy noise, robustification of the SFR model that we previously invoke, significantly improve performance of fuzzy regression model, in the case of Cauchy noise, for all scale parameters ν and number of observations n.

In Table 4, experimental results of comparison of the proposed RFR and RSFR are presented. We used fixed $p = 19$, and also the true regression vector $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0]$, $dim(\boldsymbol{b}) = 20$, while varying the scale parameter ν of the Cauchy noise imposed, and the number of observations n. It can be seen that in the case of sparse fuzzy data and imposed Cauchy noise, sparse regularization of RFR model significantly improve performance of fuzzy regression model, in the case of Cauchy noise, for all scale parameters ν and number of observations n.

In Table 5, experimental results of comparison of the proposed RFR with the baseline FR model are presented, in the case when the mixture of Cauchy and Gaussian noise was imposed. As in example in Table 1, we used fixed $p = 7$, and the true regression vector $\boldsymbol{b} = [0.4, 5, 8, 2, 3.6, 7.2, 0.2, 0.1]$, while varying the number of observations n, the scale parameter ν of the Cauchy component and the standard deviation σ of the Gaussian component in the noise mixture. It can be concluded that similar conclusions as for Table 1 hold in the case of Table 5, when the mixture is imposed.

In Table 6, experimental results of comparison of proposed SFR and RSFR are presented, in the case when the mixture of Cauchy and Gaussian noise was imposed. As in the case of Table 3, we used fixed $p = 19$, and also the true regression vector $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0]$, $dim(\boldsymbol{b}) = 20$, while varying the number of observations n, the scale parameter ν of the Cauchy component and the standard deviation σ of the Gaussian component of the mixture imposed. It can be concluded that similar conclusions as for Table 3 hold in the case of Table 5, when the mixture is imposed.

n	baseline FR			RFR		
	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$
20	0.306	0.762	1.227	0.311	0.543	0.686
100	0.636	2.056	3.558	0.010	0.126	0.158
500	0.643	1.093	3.306	0.084	0.104	0.117

Table 1: Results of the proposed RFR, in the means of mean square error E_{sqr}, in comparison to the baseline FR method, when the Cauchy noise is imposed. The true regression parameter vector is fixed on $\boldsymbol{b} = [0.4, 5, 8, 2, 3.6, 7.2, 0.2, 0.1]$. Term n represents the number of observations, while term ν represents the scale parameter for the imposed Cauchy noise.

n	baseline FR			SFR		
	$\sigma = 2$	$\sigma = 5$	$\sigma = 10$	$\sigma = 2$	$\sigma = 5$	$\sigma = 10$
20	0.223	0.284	0.536	0.069	0.161	0.328
100	0.059	0.106	0.202	0.051	0.080	0.139
500	0.047	0.053	0.073	0.039	0.037	0.036

Table 2: Results of the proposed SFR, in the means of mean square error E_{sqr}, in comparison to the baseline FR method, applied on sparse fuzzy data, when the Gaussian noise is imposed. The true regression parameter vector is fixed on $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0] \in \mathbb{R}^{p+1}$, $p = 19$. Term n represents the number of observations, while term σ represents the standard deviation for the imposed Gaussian noise.

n	SFR			RSFR		
	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$
20	0.134	0.395	0.545	0.129	0.282	0.390
100	0.665	2.261	4.290	0.064	0.115	0.161
500	0.591	2.820	5.04	0.041	0.047	0.059

Table 3: Results of the proposed RSFR and SFR methods, in the means of mean square error E_{sqr}, applied on the sparse fuzzy data, when the Cauchy noise is imposed. The true regression parameter vector is fixed on $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0] \in \mathbb{R}^{p+1}$, $p = 19$. Term n represents the number of observations, while term ν represents the scale parameter for the imposed Cauchy noise.

n	RFR			RSFR		
	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$
20	0.380	0.774	0.966	0.129	0.282	0.391
100	0.124	0.249	0.349	0.065	0.115	0.161
500	0.035	0.053	0.074	0.042	0.047	0.059

Table 4: Results of the proposed RSFR and RFR methods, in the means of mean square error E_{sqr}, applied on the sparse fuzzy data, when the Cauchy noise is imposed. The true regression parameter vector is fixed on $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0] \in \mathbb{R}^{p+1}$, $p = 19$. Term n represents the number of observations, while term ν represents the scale parameter for the imposed Cauchy noise

n	baseline FR						RFR					
	$\sigma = 5$			$\sigma = 10$			$\sigma = 5$			$\sigma = 10$		
	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$
20	0.430	0.975	1.557	0.590	1.140	1.718	0.330	0.426	0.538	0.487	0.574	0.685
100	0.283	0.659	1.063	0.407	0.730	1.114	0.242	0.332	0.405	0.360	0.448	0.526
500	0.329	0.953	1.583	0.345	0.959	1.588	0.111	0.135	0.159	0.150	0.180	0.206

Table 5: Results of the proposed RFR in comparison to the baseline FR, in the means of mean square error E_{sqr}, when the mixture of Cauchy and the Gaussian noise is imposed. The true regression parameter vector is fixed on $\boldsymbol{b} = [0.4, 5, 8, 2, 3.6, 7.2, 0.2, 0.1]$. The term n represents the number of observations, the term ν represents the scale parameter of the Cauchy component, while σ represents the standard deviation of the Gaussian component in the mixture

n	SFR						RSFR					
	$\sigma = 5$			$\sigma = 10$			$\sigma = 5$			$\sigma = 10$		
	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$	$\nu = 1$	$\nu = 3$	$\nu = 5$
20	0.776	3.944	7.335	0.791	3.936	7.298	0.100	0.148	0.210	0.165	0.218	0.287
100	0.387	1.245	2.354	0.395	1.253	2.376	0.074	0.127	0.189	0.117	0.176	0.237
500	0.364	1.194	1.880	0.349	1.175	1.870	0.045	0.067	0.084	0.060	0.081	0.099

Table 6: Results of the proposed SFR and RSFR, in the means of mean square error E_{sqr} on the sparse fuzzy data, when the mixture of Cauchy and the Gaussian noise is imposed. The true regression parameter vector is fixed on $\boldsymbol{b} = [0, 5, 2, 0, \ldots, 0] \in \mathbb{R}^{p+1}$, $p = 19$. The term n represents the number of observations, the term ν represents the scale parameter of the Cauchy component, while σ represents the standard deviation of the Gaussian component in the mixture

6. Conclusion

In this work we propose the novel Robust Fuzzy Regression model, by introducing fuzzy M-estimator that utilize Huber robust norm, where we first invoke the statistical framework in a fuzzy regression concept, by introducing additive noise that could be induced in the channel that simulate the transmission path between the fuzzy measurements and the fuzzy response. Thus we are able to handle outliers present in the fuzzy data which are modeled by the usage of heavy tailed noise. Moreover, we introduce novel Sparse Fuzzy Regression and also Robust Sparse Fuzzy Regression models, dealing with the sparsity of the actual fuzzy data in the presence of Gaussian or heavy tailed noise respectively. Experimental results obtained on synthetic data support our claims. In the future work we intend to expend the scarcity concept not only on the fuzzy regression task, but on the general task of sparse representation of fuzzy data.

REFERENCES

1. J. A. Baker: *Isometries in normed spaces.* Amer. Math. Monthly, **78** (1971), 655–658.
2. A. Marshall, I. Olkin: *Inequalities*: *Theory of Majorization and Its Applications.* Academic Press, New York, 1979.
3. H. Tanaka, S. Uegima and K. Asai, "Linear regression analysis with fuzzy model", IEEE Trans. Syst., Man, Cybern., vol. SMC-12, no. 6, pp. 903-907, Nov. 1982.
4. H. Tanaka, "Fuzzy data analysis by posibilistic linear models" , Fuzzy Sets Syst. vol. 24, pp. 363-375, 1987.
5. H. Tanaka, I. Huyashy and J. Watada, "Possibilistic linear regression analysis for fuzzy data", Eur. J. Oper. res. vol. 40. pp. 389-396, 1989.
6. H. Tanaka, H. Lee, "Interval regression by quadratic programing approach", IEEE Trans. Fuzzy Syst., vol. 6, no. 4, pp. 473-481, Nov. 1998.
7. G. Peters, "Fuzzy linear regression with fuzzy intervals", Fuzzy Sets Syst., vpl 63. pp. 45-55, 1994.
8. K. K. Yen, S. Ghoshray and G. Roig, "A linear regression model using triangular fuzzy number coefficients", Fuzzy Sets Syst. vol. 106, pp. 167-177, 1999.
9. C. Kao, C. L. Chyu, "Least-squares estimates in fuzzy regression analysis", Eur. J. Oper. Res., vol. 148, pp. 426-435, 2003.
10. C. Kao, C. L. Chyu, "A fuzzy linear regression model with better explenatory power", Fuzzy Sets Syst., vol. 126, pp. 401-409, 2002.
11. P. Diamond, "Fuzzy least squares", Inf. Sci., vol. 46., pp. 141-157, 1988.
12. P. Diamond, R. Korner, "Extended fuzzy linear models and least squares estimates", Commput. Math. Appl., vol. 33, pp. 15-32, 1997.
13. M. Ma, M. Friedman and A. Kandel, "General fuzzy least squares", Fuzzy Sets Syst., vol. 88, pp. 107-118, 1997.

14. B. Wu, N. F. Tseng, "A new approach to fuzzy regression models with application to business cycle analysis", Fuzzy Sets Syst., vol. 130, pp. 33-42, 2002.

15. D. Zhang, L. F. Deng, K. Y. Cai and A. So, "Fuzzy nonlinear regression with fuzzified radial basis function network", IEEE Trans. Fuzzy Syst., vol. 13, no. 6, pp. 742-760, Dec. 2005.

16. B. Kim and B. B. Bishu, "Evaluation of fuzzy linear regression models by comparing membership functions", Fuzzy Sets Syst., vol. 100, pp. 343-352, 1998.

17. L. H. Chen, C. C. Hsueh, "Fuzzy regression models using least-squares method based on the concept of distance", IEEE Trans. on Fuzzy Systems, vol. 17, no. 6, pp. 1259-1272, December 2009.

18. R. Tibshirani, "Regression shrinkage and selection via the LASSO", J. R. Statist. Soc. B, vol. 58, no. 1, (1996), pp. 267-288.

19. R. Tibshirani, Michael Saunders, "Sparsity and smoothness via the fuzed LASSO", J. R. Statist. Soc. B, vol. 58, no. 1, (2005), 67, Part 1, pp. 91-108.

20. S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation", IEEE Trans. Signal Process., vol. 57, no. 7, pp. 2479-2493, 2009.

21. S. M. Kay, "Fundamentals of statistical signal processing: estimation theory", Prentice Hall PTR, Upper Saddle River, NJ, 1993.

22. L. H. Chen and C. C. Hsueh, "Mathematical programming method for formulating fuzzy regression model based on distance criterion", IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 37, no. 3, pp. 705-712, Jun. 2007.

23. G.J. Klir, B. Yuan, "Fuzzy sets and fuzzy logic, theory and applications", Prentice Hall, P.T.R, Upper Sadle River, New Jersey 07458, 1995.

24. D. J. Olive, "Applied robust statistic", Course notes, Southern Illinois University, Department of Mathematics, 23. June, 2008.

25. R. Andersen, "Modern methods for robust regression, quantitative applications in the social sciences", 152. Los angeles, CA: Sage Publications, 2008.

26. P. J. Huber, Elvezio M. Ronchetti, "Robust statistics (2nd ed.)" Hoboken, NJ: John Wiley and Sons Inc., 2009.

27. J. Nocedal, S. J. Wright, "Numerical optimization", Springer-Verlag, New York, Inc., 1999.

28. S. P. Boyd, L. Vanednberghe, "Convex optimization", Cambrige UP, 1990.

29. M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing", Proc. IEEE, vol. 98, no. 6, pp. 972-982, 2010.

30. R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling", Proc. IEEE, vol. 98, no. 6, pp. 1045-1057, 2010.

31. R. Obradovic, M. Janev, B. Antic, V. Crnojevic, N. I. Petrovic, "Robust sparse image denoising", 18th IEEE International Conference on Image Processing (ICIP), pp. 2569-2572, 2011.

32. R. Garnett, T. Huegerich, C. Chui, "A universal noise removal algorithm with impulse detector", IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1747-1754, November 2005.

33. S. Shulte, V. De Witte, M. Nachtegael, D. Van der Weken, and E. E. Kerre, "Fuzzy random impulse noise reduction method", Fuzzy Sets and Systems, vol. 158, pp. 270-283, 2007.

34. S. Perkins, k. Lacker, J. Theiler, "Grafting: fast, incremental feature selection by gradient descent in function space", Journal of Machine Learning Research 3 (2003) 1333-1356.

35. S. Perkins, J. Theiler, "Online feature Selection using grafting", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

36. E. Lughofer, S. Kindermann, M. Pratama, J. de J. Rubio, "Top–down sparse fuzzy regression modeling from data with improved coverage", Int. J. Fuzzy Syst. (2017) 19: 1645.

37. M. Luo, F. Sun, H. Liu, "Sparse fuzzy c-regression models with application to T-S fuzzy systems identification", 2014 IEEE International Conference on Fuzzy Systems, 6-11 July, 2014.

38. T. Dam, A. Deb, "Block sparse representations in modified fuzzy c-regression model clustering algorithm for TS fuzzy model identification", 2015 IEEE Symposium Series on Computational Intelligence, 7-10 Dec. 2015.

39. S. Yeylaghia M. Otadib N. Imankhan, "A new fuzzy regression model based on interval-valued fuzzy neural network and its applications to management", Beni-Suef University Journal of Basic and Applied Sciences, 6 (2017) 106–111

40. A. F. R. L. de Hierro ; J. Martinez-Moreno ; C. Aguilar-Pena ; C. R. L. de Hierro, "Estimation of a fuzzy regression model using fuzzy distances", IEEE Transactions on Fuzzy Systems, vol. 24 (2016).

41. H.S. Wang, G.D. Li, G.H. Jiang, "Robust regression shrinkage and con- sistent variable selection through the LAD-Lasso", Journal of Business and Economic Statistics, 25 (2007), 347-355

42. X. Gao, J. Huang, "Asymptotic analysis of high-dimensional LAD regression with Lasso", Statistica Sinica, 20 (2010), 1485-1506.

43. J. Mairal, G. Sapiro, M. Elad, "Learning multiscale sparse representations for image and video restoration", Multiscale Model. Simul., 7(1), 214–241.

44. N. Ouzir, A.Basarab, H. Liebgott, B. Harbaoui, J. Tourneret, "Motion estimation in echocardiography using sparse representation and dictionary learning", IEEE Transactions on Image Processing, vol. 27 (2018)

45. X. Yang W. Wu, K. Liu, W. Chen, P. Zhang, Z. Zhou, "Multi-sensor image super-resolution with fuzzy cluster by using multi-scale and multi-view sparse coding for infrared image", Multimed Tools Appl (2017) 76: 24871.

46. L. Lv, D. Zhao, Q. Deng, "Image clustering based on deep sparse representations", 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 6-9 Dec. 2016

47. L. Li, W. Yao, "Fully Bayesian logistic regression with hyper-Lasso priors for high-dimensional feature selection", arXiv:1405.3319

48. A. Boulesteix, R. De Bin, X. Jiang, M. Fuchs, "IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data", Computational and Mathematical Methods in Medicine, vol. 2017 (2017), Article ID 7691937, doi.org/10.1155/2017/7691937

49. H.J. Zimmermann, "Fuzzy set theory-and its applications", Springer Seience and Business Media, New York, 2001.

(Received 27.12.2017)
(Revised 12.09.2019)

Danilo Rapaić
University of Novi Sad,
Faculty of Technical Sciences,
Trg D. Obradovića 6,
21000 Novi Sad,
Serbia
E-mail: *rapaicd@yahoo.com*

Lidija Krstanović
University of Novi Sad,
Faculty of Technical Sciences,
Department of Fundamentals Sciences,
Chair of Engineering Animation ,
Trg D. Obradovića 6,
21000 Novi Sad,
Serbia
E-mail: *lidijakrstanovic@uns.ac.rs*

Nebojša Ralević
University of Novi Sad,
Faculty of Technical Sciences,
Trg D. Obradovića 6,
21000 Novi Sad,
Serbia
E-mail: *nralevic@uns.ac.rs*

Ratko Obradović
University of Novi Sad,
Faculty of Technical Sciences,
Department of Fundamentals Sciences,
Chair of Engineering Animation ,
Trg D. Obradovića 6,
21000 Novi Sad,
Serbia
E-mail: *obrad_r@uns.ac.rs*

Djuro Klipa
University of Novi Sad,
Faculty of Technical Sciences,
Trg D. Obradovića 6,
21000 Novi Sad,
Serbia
E-mail: *djklipa@minrzs.gov.rs*