STUDIA UNIV. "BABEŞ–BOLYAI", INFORMATICA, Volume **XLIII**, Number 1, 1998

A NEW REGRESSION TECHNIQUE BASED ON FUZZY SETS

HORIA F. POP

Abstract. The purpose of this paper is to present a new regression algorithm based on fuzzy sets and to describe some of its interesting properties. As shown in [11], the application of this algorithm and other conventional ordinary and weighted least squares and robust regression methods to relevant data sets proves that the performance of the procedure described in this paper exceeds that of the ordinary least squares method and equals, and often exceeds that of weighted or robust methods, including the two fuzzy methods proposed in [9] and [7]. Based on our previous experience, we introduce two new regression line quality criteria. In [11] we emphasized the effectiveness and the generality of these criteria for diagnosing the linearity of calibration lines.

Calibration of the instrumental response is a fundamental requirement for all instrumental analysis techniques. In statistical terms, a calibration refers to the establishment of a predictive relation between the controlled or independent variable (e.g. the concentration of a standard) and the instrumental response. The common approach to this problem is to use the linear least squares method. The ordinary least squares regression (LS) is based on the assumption of an independent and normal errors distribution with uniform variance (homeoscedastic). Much more common in practice, however, are the heteroscedastic results, where the y-direction error is concentration-dependent and/or the presence of outliers.

In practice the actual shape of the error distribution function and its variance are usually unknown, so we must investigate the consequences if the conditions stated above are not met. Generally, the least squares method does not lead to the maximum likelihood estimate. Despite the fact that the least squares method is not optimal, there is a justification for using it in the cases where the conditions are only approximately met.

In particular, the Gauss-Markov theorem states that if the errors are random and uncorrelated, the least squares method gives the best linear unbiased estimate of the parameters, meaning that out of all the functions for which each parameter is a linear function, the least squares method is that for which the variances of the parameters are the smallest [8].

1991 *Mathematics Subject Classification.* 68H30, 68T10.

1991 *CR Categories and Descriptors.* I.5.1 [**Pattern Recognition**]: Models – *Fuzzy set*; I.5.3 [**Pattern Recognition**]: Clustering – *Algorithms, Similarity measures.*

Nevertheless, if the tails of the experimental error distribution contain a substantially larger proportion of the total area than the tails of a Gaussian distribution, the "best linear" estimate may not be very good, and there will usually be a procedure in which the parameters are not linear functions of the data, that gives lower variances for the parameter estimates than the least squares method does, that is the robust and resistant method.

estimates with error effect of

given by a not robust.

to the applies discrepant with There are error being data points modeled) to some experiment

determining what the least squares the least weakness of the of residuals, ignored. For connect the a smooth or

The purpose of the present paper is to introduce a new regression technique based on fuzzy sets. We will show how the well-known Fuzzy n-Lines algorithm may be modified in order to be used to produce a single fuzzy set. We will also study the interesting properties of the fuzzy set produced by running the algorithm.

1. The Fuzzy n-Lines algorithm

Let us consider a data set $X = \{x^1, \ldots, x^p\} \subset \mathbb{R}^s$. Let us suppose that the cluster substructure of X corresponds to linear clusters. The task is to generate a fuzzy partition $P = \{A_1, \ldots, A_n\}$ of X corresponding to its cluster substructure. We are also interested in obtaining a geometrical characterization of the detected clusters [2, 3].

In what follows we will suppose that each fuzzy set A_i is represented by a linear prototype L_i which passes through the point V^i and has the direction of the unit vector u^i. We will denote this line with

$L_i(v^i, u^i)$:

$$L_i(v^i, u^i) = \{y \in \mathbb{R}^d | y = v^i + tu^i, t \in \mathbb{R}\}.$$

Let us consider the scalar product in $\mathbb{R}^s$ given by

$$\langle x, y \rangle = x^T M y,$$

where $M \in \mathbf{M}_s(\mathbb{R})$ is a symmetrical and positively defined matrix. If M is the unit matrix, then the scalar product is the usual one, $\langle x, y \rangle = x^T y$.

Let us consider the norm in $\mathbb{R}^s$ induced by the scalar product,

$$\|x\| = \langle x, x \rangle^{1/2}, \text{ for every } x \in \mathbb{R}^s,$$

and the distance d in $\mathbb{R}^s$ induced by this norm,

$$d(x, y) = \|x - y\|, \text{ for every } x, y \in \mathbb{R}^s.$$

The distance between the point x and the line L_i is

$$d(x, L_i) = \min_{y \in L_i} d(x, y) = \left(\|x - v^i\|^2 - \langle x - v^i, u^i\rangle^2\right)^{1/2}.$$

The local metric d_i induced by d and the class A_i is

$$d_i(x, L_i) = AI(x)d(x, L_i).$$

The dissimilarity between a point x^j and the class A_i is chosen as being the square of the local distance from the point to the prototype line:

$$D(x^j, L_i) = d_i^2(x^j, L_i) = \left(A_i(x^j)\right)^2 \left(\|x - v^i\|^2 - \langle x - v^i, u^i\rangle^2\right).$$

The inadequacy $I(A_i, L_i)$ between the class A_i and its prototype L_i is defined as

$$I(A_i, L_i) = \sum_{j=1}^{p} D(x^j, L_i).$$

Let $L = \{L_1, \ldots, L_n\}$ be the representation of the fuzzy partition P. The inadequacy between P and L is written as

$$J(P, L) = \sum_{i=1}^{n} I(A_i, L_i)$$

that is

$$J(P, L) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(A_i(x^j)\right)^2 \left(\|x - v^i\|^2 - \langle x - v^i, u^i\rangle^2\right). \tag{1}$$

We are interested to find the fuzzy partition P and its representation P that minimize the criterion function J defined through relation (1).

Since a global minima of this problem can not be reached, we will give an approximative method for determining a local minima [2, 3].

Theorem 1.1. *The fuzzy partition $P = \{A_1, \ldots, A_n\}$ of X which minimizes the function $J(\cdot, L)$ given by relation (1) is characterized by*

$$A_i(x^j) = \frac{1}{\sum_{k=1}^{n} \frac{d^2(x^j, L_i)}{d^2(x^j, L_k)}}, is \forall i, d(x^j, L_i) \neq 0 \tag{2}$$

and, respectively, if for a certain x^j there is at least an L_i so that $d(x^j, L_i) = 0$, the memberships of x^j fulfill the conditiopn

$$A_i(x^j) = 0, \forall i \text{ so that } d(x^j, L_i) \neq 0.$$

Theorem 1.2. *The set of prototypes $L = \{L_1, \ldots, L_n\}$ that minimizes the function $J(P, \cdot)$ given by relation (1) is characterized by*

$$v^i = \frac{\sum_{j=1}^{p} (A_i(x^j))^2 x^j}{\sum_{j=1}^{p} (A_i(x^j))^2}; \tag{3}$$

u^i is the unit eigenvector corresponding to the largest eigenvalue of the matrix S_i defined as

$$S_i = M \left(\sum_{j=1}^{p} (A_i(x^j))^2 (x^j - v^i)(x^j - v^i)^T \right) M. \tag{4}$$

For the proof of these two theorems please see [2].

Following these two results, the linear clusters structure of the data set X may be built using the following **Fuzzy n-Lines algorithm**:

S1.: Chose an arbitrary partition $P^{(0)} = \{A_1, \ldots, A_n\}$ of X and set $l = 0$.

S2.: Computes the prototypes $L_i(v^i, u^i)$, $i = 1, \ldots, n$ of the partition $P^{(l)}$ using (2) and (3).

S3.: Determines a new partition $P^{(l+1)}$ using relation (1).

S4.: If partitions $P^{(l)}$ and $P^{(l+1)}$ are close enough, that is if

$$\|P^{(l+1)} - P^{(l)}\| < \epsilon,$$

where ϵ is a predefined value, then **stop**, else increase l by 1 and continue from step **S2**.

2. The Fuzzy 1-Lines algorithm

Let us consider a data set $X = \{x^1, \ldots, x^p\} \subset \mathbb{R}^s$. Let us suppose that the set X does not have a clear clustering structure or that the clustering structure corresponds to a single fuzzy set. Let us admit that this fuzzy set, denoted by A, may be characterized by a linear prototype, denoted by $L = (v, u)$, where v is the center of the class and u, with $\|u\| = 1$, is its main dirrection. We rise the problem of finding the fuzzy set that represents the most suitable the given data set. We propose ourselves to do this by minimizing a criterion function similar to those presented in [6, 10, 5, 2, 3].

In order to obtain the criterion function we will have in mind that we wish to determine a fuzzy partition $\{A, \overline{A}\}$. The fuzzy set A is characterized by the prototype L. In what it concerns the complementary fuzzy set, $\overline{A}$, we will consider that the dissimilarity between its hypotetical protptype and the points x^j is constant and equal to $\frac{\alpha}{1-\alpha}$, where α is a constant from $(0, 1)$, with a role to be seen later in this paper.

Based on the notations from the previous section, the inadequacy between the fuzzy set A and its prototype L will be

$$I(A,L)=\sum_{j=1}^{p}(A(x^j,L))^2D(x^j,L),$$

and the inadequacy between the complementary fuzzy set $\overline{A}$ and its hypothetical prototype will be

$$\sum_{j=1}^{p}(\overline{A}(x^j))^2\cdot k.$$

Thus, the criterion function $J:F(X)\times\mathbb{R}^d\to\mathbb{R}^+$ becomes

$$J(A,L;\alpha)=\sum_{j=1}^{p}(A(x^j))^2\cdot d^2(x^j,L)+\sum_{j=1}^{p}(\overline{A}(x^j))^2\cdot\frac{\alpha}{1-\alpha},$$

where $\alpha\in(0,1)$ is a fixed constant.

With respect to the minimization of the criterion function J the following two results are valid.

Theorem 2.1. *Let us consider the fuzzy set A of X. The prototype $L=(v,u)$ that minimizes the function $J(A,\cdot)$ is given by*

$$v=\frac{\sum_{j=1}^{p}(A(x^j))^2x^j}{\sum_{j=1}^{p}(A(x^j))^2}; \tag{5}$$

u is the eigenvector corresponding to the maximal eigenvalue of the matrix

$$S=M\sum_{j=1}^{p}(A(x^j))^2(x^j-v)(x^j-v)^TM. \tag{6}$$

Theorem 2.2. *Let us consider a certain prototype L. The fuzzy set A that minimizes the function $J(\cdot,L)$ is given by*

$$A(x^j)=\frac{\frac{\alpha}{1-\alpha}}{\frac{\alpha}{1-\alpha}+d^2(x^j,L)}. \tag{7}$$

The optimal fuzzy set will be determined by using an iterative method where J is succesively minimized with respect to A and L. The proposed algorithm will be called **Fuzzy 1-Lines**:

S1: We choose the constant $\alpha \in [0,1]$. We initialize $A(x) = 1$, for every $x \in X$, and $l = 0$. Let us denote by $A^{(l)}$ the fuzzy set A determined at the l-th iteration.

S2: We compute the prototype $L = (v,u)$ of the fuzzy set $A^{(l)}$ using the relations (5) and (6).

S3: We determine the new fuzzy set $A^{(l+1)}$ using the relation (7).

S4: If the fuzzy sets $A^{(l+1)}$ and $A^{(l)}$ are closed enough, i.e. if

$$\|A^{(l+1)} - A^{(l)}\| < \epsilon,$$

where ϵ has a predefined value, then **stop**,
else increase l by 1 and go to step **S2**.

We experimentally noticed that a good value of ϵ with respect to the similarity of $A^{(l)}$ and $A^{(l+1)}$ would be $\epsilon = 10^{-5}$. We used this value for all the computations performed in this paper.

In order to avoid the dependency of the memberships of the scale, in practice we will use in the relation (7), instead of the distance d the normalized distance d_r given by

$$d_r(x^j, L) = \frac{d(x^j, L)}{\max_{x \in X} d(x, L)}.$$

Thus, the relation used to determine the memberships is

$$A(x^j) = \frac{\frac{\alpha}{1-\alpha}}{\frac{\alpha}{1-\alpha} + d_r^2(x^j, L)}. \tag{8}$$

The algorithm modified in this way will be called **Modified Fuzzy 1-Lines**:

S1: We chose the constant $\alpha \in [0,1]$. We initialize $A(x) = 1$, for every $x \in X$, and $l = 0$. Let us denote by $A^{(l)}$ the fuzzy set A determined at the l-th iteration.

S2: We compute the prototype $L = (v,u)$ of the fuzzy set $A^{(l)}$ using the relations (5) and (6).

S3: We determine the new fuzzy set $A^{(l+1)}$ using the relation (8).

S4: If the fuzzy sets $A^{(l+1)}$ and $A^{(l)}$ are closed enough, i.e. if

$$\|A^{(l+1)} - A^{(l)}\| < \epsilon,$$

where ϵ has a predefined value (10^{-5}), then **stop**, else increase l by 1 and go to step **S2**.

3. The properties of the produced fuzzy set

In this section we will study the properties of the fuzzy set obtained via the algorithm presented above.

Theorem 3.1. *Let X be a given data set and let A and L be the fuzzy set respectively its propotype as they were determined by the Modified Fuzzy* 1*-Lines algorithm. The following relations are valid (the notations are those used by now):*

(i): $A(x) = 1 \iff d(x, L) = 0$;
(ii): $A(x) = \alpha \iff d_r(x, L) = 1$;
(iii): $A(x) \in [\alpha, 1]$ *for every* $x \in X$;
(iv): $\alpha = 0 \iff A(x) = 0$ *for every* $x \in X$;
(v): $\alpha = 1 \iff A(x) = 1$ *for every* $x \in X$;
(vi): $A(x^i) < A(x^j) \iff d(x^j, L) < d(x^i, L)$;
(vii): $A(x^i) = A(x^j) \iff d(x^j, L) = d(x^i, L)$.

The proof is quite simple. All the relations come from 7 □.

The algorithm presented here converges towards a local minimum. Normally, the results of the algorithms of this type are influenced by the initial partition considered [4, 1]. In this case the initial fuzzy set considered being X, the obtained optimal fuzzy set is the one situated in the vicinity of X, and this makes the algorithm even more attractive.

Let us remark that the role of the constant α is to affect the polarization of the partition $\{A, \overline{A}\}$. Also, now is clear why α was chosen to be in $(0, 1)$ and the values 0 and 1 were avoided.

By using the relative dissimilarities D_r, this method is independent of the linear transformations of the space.

Having in mind the properties (i) – (vii) and the remarks above, the fuzzy set A determined here may be called **fuzzy set associated to the classical set** X **and to the membership threshold** α.

On the model of the theory presented above we may build a theory to determine the fuzzy set A represented by a point proptotype, or by any geometrical prototype.

As seen above, the (Modified) Fuzzy 1-Lines algorithm produces the fuzzy set associated to the classical set X and to the membership threshold α, together with its linear representation. This particular procedure may also be called **Fuzzy Regression**.

4. Quality indices

Taking into account the contradictory values of different quality coefficients and the diversity of methods concerning their algorithm, we are introducing two new quality coefficients, namely Q_1 and Q_2.

The first coefficient, Q_1 refers to the maximum of absolute residuals,

$$Q_1 = \sqrt{\sum_{j=1}^{p} \left(\frac{r_j}{\max |r_j|} \right)^2}, \tag{9}$$

and Q_2 is referring to the mean of absolute residuals,

$$Q_2 = \sqrt{\sum_{j=1}^{p} \left(\frac{r_j}{\overline{r}}\right)^2}, \tag{10}$$

where r_j is the distance from the point x^j to the regression line L,

$$r_j = d(x^j, L)$$

.

Theorem 4.1. *The following relations are valid:*

(i): $1 \leq Q_1 \leq \sqrt{p}$;
(ii): $\sqrt{p} \leq Q_2 \leq p$.

The proof is straightforward □.

Based on the result above we will introduce the normalized variants of these coefficients, namely NQ_1 and NQ_2, which take values within the range $[0, 1]$, and thus appear to be more practical:

$$NQ_1 = \frac{Q_1 - 1}{\sqrt{p} - 1} \tag{11}$$

and

$$NQ_2 = \frac{Q_2 - \sqrt{p}}{p - \sqrt{p}}. \tag{12}$$

Based on our practical experiments [11], it may be stated that these new quality coeficients concerning the goodness of fit proposed in this paper confirm our main conclusions and are in a good agreement with the statements in the analytical literature.

Conclusions

A new fuzzy regression algorithm has been described in this paper. It was compared with conventional ordinary and weighted least squares and robust regression methods. The application of these different methods to relevant data sets proved that the performance of the procedure described in this paper exceeds that of the ordinary least squares method and equals, and often exceeds that of weighted or robust methods, including the two fuzzy methods proposed in [9] and [7].

Moreover, we underline the effectiveness and the generality of
the two new criteria proposed in this paper for diagnosing the linearity of calibration lines in analytical chemistry.

In addition, we have to emphasize that the fuzzy regression method discussed above includes not only the estimates of the parameters, but also additional information in the form of membership degrees. In this way the fuzzy regression algorithm gives a new aspect to regression methods. Using the membership degrees of each point to the regression line we may compute the informational energy and/or the informational entropy. These quantities allow us to appreciate the presence of outliers and the linearity of the regression line.

References

[1] Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Functions Algorithms*. Plenum Press, New York, (1981).

[2] Bezdek, J. C., Coray, C., Gunderson, R., and Watson, J. Detection and characterisation of cluster substructure: I. Linear structure: Fuzzy C-lines. *SIAM Journal on Applied Mathematics 40*, 2 (1981), 339–357.

[3] Bezdek, J. C., Coray, C., Gunderson, R., and Watson, J. Detection and characterisation of cluster substructure: II. Linear structure: Fuzzy C-varieties and convex combinations thereof. *SIAM Journal on Applied Mathematics 40*, 2 (1981), 358–372.

[4] Dumitrescu, D. *Classification Theory (Romanian)*. "Babes-Bolyai" University Press, Cluj-Napoca, (1991).

[5] Dumitrescu, D., Pop, H. F., and Sârbu, C. Fuzzy hierarchical cross-classification of Greek muds. *Journal of Chemical Information and Computer Sciences 35* (1995), 851–857.

[6] Dumitrescu, D., Sârbu, C., and Pop, H. F. A fuzzy divisive hierarchical clustering algorithm for the optimal choice of set of solvent systems. *Analytical Letters 27*, 5 (1994), 1031–1054.

[7] Hu, Y., Smeyers-Verbeke, J. and Massart, D.L. An algorithm for fuzzy linear calibration. *Chemometrics and Intelligent Laboratory Systems 8* (1990), 143–155.

[8] Klimov, G. *Probability Theory and Mathematical Statistics*, Mir Publishers, Moscow (1986).

[9] Otto, M., and Bandemer, H. Calibration with imprecise signals and concentrations based on fuzzy theory. *Chemometrics and Intelligent Laboratory Systems 1* (1986), 71–78.

[10] Pop, H. F., Dumitrescu, D., and Sârbu, C. A study of Roman pottery (terra sigillata) using hierarchical fuzzy clustering. *Analitica Chimica Acta 310* (1995), 269–279.

[11] Pop, H. F., and Sârbu, C. A new fuzzy regression algorithm. *Journal of Analytical Chemistry 68* (1996), 771–778.

"Babes-Bolyai" University, Faculty of Mathematics and Computer Science, RO-3400 Cluj-Napoca, Romania

E-mail address: hfpop@cs.ubbcluj.ro