Philosophical Interpretation of Connection of Robust Statistics and Fuzzy Logic: The Robust Fuzzy Clustering

Vladimir Djordjevic[1*], Vojislav Filipovic[1]
[1]Department for Automatic Control, Robotics and Fluid Technique, Faculty of Mechanical and Civil Engineering, University of Kragujevac, Kraljevo (Serbia)

Clustering methods have the key role in pattern recognition, computer vision, and control. In real applications, the data are corrupted with stochastic noise which often has outliers. It follows that clustering techniques need to be robust. It is observed that robust statistics and fuzzy set theory have much in common. Namely, the concept of weight functions in robust statistics can be related to the concept of membership function in fuzzy set theory. In the paper proposed the new objective function for cluster analysis. For the clustering the modified Gustafson-Kessel algorithm is used and the modification is based on possibility theory. The final goal is membership function determination. That is the important part of the Takagi–Sugeno models which represent the fuzzy model of nonlinear dynamic systems.

Keywords: Robust statistics, IRLS, Possibility function, Robust clustering

1. INTRODUCTION

Robustness is a key attribute in engineering systems. It means that the performance of algorithm (in identification, estimation, and control) should not be affected significantly by small deviations from the assumed model. Also, it is important that it should not deteriorate drastically due to noise and outliers. In this paper, we consider clustering techniques which are used in different fields

(i) Pattern recognition [1]
(ii) Computer and robot vision [2]
(iii) Control [3]

In real application data have outliers [4] and different procedures have to be robust. Two disciplines, robust statistics and fuzzy logic have developed independently. But, as we will see in this paper, they have much in common. That explains the claim of proponents of fuzzy set theory that a fuzzy approach is more tolerant to model variations and disturbances in comparison with the crisp approach.

In this paper, we first introduce some concept from robust statistics [5-6] (min-max property, infinite signal function and breakdown paints). After that, we will describe, shortly, a few concepts from fuzzy logic. Finally, it will be established some philosophical connection between both areas.

The main goal of the paper is to find robust cluster procedure with small sensitivity to outliers. The unified view of the problem is presented in [7]. In this paper, we consider robust version of Gustafson-Kessel (GK) algorithm [8]. That algorithm is not considered in [7]. Also, we consider a possibilistic approach to clustering. A heuristic version of that theory is described in [10]. The main result is possibilistic Gustafson-Kessel algorithm (PGK). The primary objective of the possibilistic approach is to achieve membership value that is possibilistic, i.e. the membership value of a point in a class represents the possibility of the point belonging to the class.

For the robustness the number of clusters is also important. That problem is not considered in this paper but attractive solution is given in [11]. Here is of interest cluster validity. Cluster validity measures the correctness of partition generated by a clustering algorithm. In [12] as a model validity test is used Kolmogorov-Smirnov test.

2. ROBUST STATISTICS

Robustness is a very important notion is modern science. In statistics, the robustness is low sensitivity to distribution changes of real processes. At present, in statistical sense, there are two key approaches to robustness

(i) Quantitative robustness [13] known as a Huber's minmax approach
(ii) Qualitative robustness [14] which is based on the concept of influence function.

We will first consider Huber's approach. Usually, the problem of parameter estimation is based on the assumption that the stochastic disturbance has a Gaussian distribution. Practical studies [4] show that in a population of observations there are rare large observations (outliers) and the result is that stochastic disturbance has a non-Gaussian disturbance. Such case is considered in [13] where class of distributions is modeled as

$$P_\varepsilon = \left\{ P : P = (1-\varepsilon)N + \varepsilon G, \quad G \text{ is symetric} \right\} \qquad (1)$$

where $\varepsilon \in [0,1)$ is the contamination degree and $N(0,\sigma^2)$ denotes a zero-mean Gaussian distribution with a variance σ^2. Applying Huber's methodology [5] and [13] the least favorable probability density on a class (1) is obtained

$$p^*(e(k)) = \begin{cases} \dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \exp\left\{ -\dfrac{(e(k))^2}{2\sigma^2} \right\}, & |e(k)| \le k_\varepsilon \\ \dfrac{1-\varepsilon}{\sqrt{2\pi}\sigma} \exp\left\{ -\dfrac{k_\varepsilon}{\sigma^2}\left(|e(k)| - \dfrac{k_\varepsilon}{2} \right) \right\}, & |e(k)| > k_\varepsilon \end{cases} \qquad (2)$$

*Corresponding author: Dositejeva 19, Kraljevo (Serbia), djordjevic.v@mfkv.kg.ac.rs

where $e(k)$ is a stochastic disturbance and where the relationship between the contamination degree ε and the Huber's parameter k_ε is given in the following relation

$$\frac{2\Phi_N(k_\varepsilon)}{k_\varepsilon} - 2\Phi_N(-k_\varepsilon) = \frac{\varepsilon}{1-\varepsilon},\ \Phi_N = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \quad (3)$$

The good performance of parameter estimation algorithm is provided for $k_\varepsilon \in [2,4]$. The best performance is accomplished for $k_\varepsilon = 3$ as is shown in [14-17]. In cited references, the robust recursive identification of MIMO (multiple-input multiple-output) is considered.

In what follows we will consider estimation of location and scale parameters.

Remark 1. It is possible to describe outliers with other probability distribution in comparison with (1). That is heavy-tailed (fat-tailed) distributions [18] whose density tails tend to zero more slowly than the normal density. An example is the Cauchy distribution with density

$$f(x) = \frac{1}{\pi(1+x^2)} \quad (4)$$

It is a particular case of the Student (or t) densities with $\nu > 0$ degrees of freedom

$$f_\nu(x) = c_\nu \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (5)$$

where c_ν is a constant

$$c_\nu = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \quad (6)$$

and $\Gamma(\cdot)$ is the gamma function. This family contains all degrees of heavy-tailedness. When $\nu \to \infty$, f_ν tends to the standard normal density. For $\nu = 1$ we have the Cauchy distribution.

We now consider M-estimate of location parameter. Using relation (2) we can define the function

$$\Phi(x) = -\log p^*(x) \quad (7)$$

or explicitly

$$\Phi(x) = \begin{cases} \frac{x^2}{2\sigma^2} + \ln\frac{\sqrt{2\pi}\sigma}{1-\varepsilon}, & |x| \le k_\varepsilon \\ \frac{k_\varepsilon}{\sigma^2}\left(|x| - \frac{k_\varepsilon}{2}\right) + \ln\frac{\sqrt{2\pi}\sigma}{1-\varepsilon}, & |x| > k_\varepsilon \end{cases} \quad (8)$$

Let us define the model for location

$$y(k) = \theta + e(k) \quad (9)$$

where $y(k)$ are the measurement, θ is the location parameter and $e(k)$ is the stochastic non-Gaussian process. According to maximum likelihood methodology criterion for parameter θ estimation in (9) is

$$J(\theta) = E\{\Phi(y(k) - \theta)\} \quad (10)$$

where $E\{\cdot\}$ is expectation operator. It is possible to approximate the last relation with empirical functional

$$J_k(\hat{\theta}) = \frac{1}{k}\sum_{i=1}^{k} \Phi(y(k) - \hat{\theta}(k)) \quad (11)$$

where

$$\hat{\theta}(k) = \arg\min_\theta \sum_{i=1}^{k} \Phi(y(k) - \theta) \quad (12)$$

If $\Phi(\cdot)$ is differentiable, differentiating with respect to θ yields

$$\sum_{i=1}^{k} \Psi(y(i) - \hat{\theta}(i)) = 0 \quad (13)$$

with $\Psi(\cdot) = \Phi'(\cdot)$. For probability distribution (1) the function $\Psi(\cdot)$ is known as a Huber's function and has the analytical expression

$$\Psi(x) = \begin{cases} x, & |x| \le k_\varepsilon \\ k_\varepsilon, & |x| > k_\varepsilon \end{cases} \quad (14)$$

and is showed in the next figure.

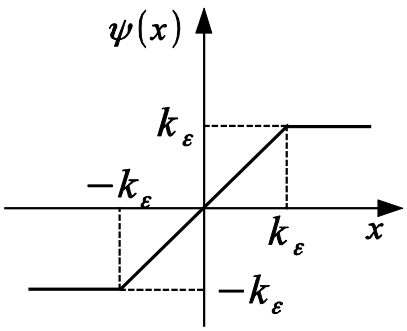

Figure 1: Huber's function

Let us notice that solution of equation (13) is estimated value of location parameter.

The same story is valid for estimation of scale parameter. In this case, the loss function is

$$\Phi\left(\frac{x}{\sigma}\right) \quad (15)$$

where σ is scale parameter.

The qualitative robustness is presented in [6] and has two concepts:

(i) Influence function which is an infinitesimal approach (small deviations from the model assumptions should impair the performance of estimation only by a small amount). That concept has a local character;

(ii) Breakdown point which means that larger deviations from the model assumptions have not catastrophic consequences. This concept has a global character.

The influence function can be defined using theory of a von Mises functionals [19].

Definition 1. [6] The influence function of a von Mises functional T on probability distribution F is given by

$$IF\left(x,T,F\right)=\lim_{\varepsilon\to 0}\frac{T\left[\left(1-\varepsilon\right)F+\varepsilon\Delta_x\right]-T\left(F\right)}{\varepsilon} \quad (16)$$

whrere Δ_x is the probability measure which puts mass 1 at the point x.

The breakdown point is defined in [20]. Let us suppose that Z consist of N data points and T is estiomator. Also, Z' is set which is given by replacing any M of the original data points by arbitrary values. Let us denote with bias (M,T,Z) the maximum bias in the estimate caused by such a contamination

$$\text{bias}\left(M,T,Z\right)=\sup_{Z'}\left\|T\left(Z'\right)-T\left(Z\right)\right\| \quad (17)$$

If the bias is infinite, the M outliers have an arbitrary large effect on T and, thus, the estimator break down. The definition of breakdown point is

$$\varepsilon_{BP}^{*}=\min\left\{\frac{M}{N}:\text{bias}\left(M,T,Z\right)\text{ is infinite}\right\} \quad (18)$$

Finally, we will consider important, from the computation point of view, concept known as the iteratively reweighted least square (IRLS) [21-23]. Let us consider relation (13) without index of estimate of location parameter θ. That relation we can rewrite in the next form

$$\sum_{i=1}^{k}\left(y\left(i\right)-\theta\right)w\left(y\left(i\right)-\theta\right)=0 \quad (19)$$

whereby w is weight coefficient

$$w\left(y\left(i\right)-\theta\right)=\frac{\Psi\left(y\left(i\right)-\theta\right)}{y\left(i\right)-\theta} \quad (20)$$

From relation (19) one can get

$$\theta=\frac{\sum_{i=1}^{k}w\left(y\left(i\right)-\theta\right)y\left(i\right)}{\sum_{i=1}^{k}w\left(y\left(i\right)-\theta\right)} \quad (21)$$

It follows that θ is a weighted mean of the $y\left(i\right)$ and can be solved iteratively.

It is important to note that for complex dynamic systems it is possible to approximate and relax computation by using principal component analysis [24].

The behaviour of weight w for probability distribution model (1) and Huber's function (14) is presented in the next figure.

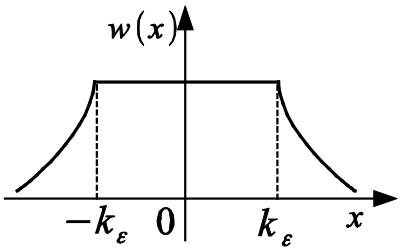

Figure 2: The weight function in location parameter esitmation

3. FUZZY LOGIC

Fuzzy logic now has high theoretical level with applications in many fields. The key notation in fuzzy logic is the set membership function which is used for characterization of fuzzy sets [25].

A fuzzy set A on the universe set X is a set defined by a membership function $\mu_A\left(x\right)$ represent a mapping

$$\mu_A\left(x\right):X\to\left[0,1\right] \quad (22)$$

Here the value of $\mu_A\left(x\right)$ for the fuzzy set A is called the membership value or the grade of membership of $x\in X$. The membership value means the degree of x belonging to the fuzzy set A.

Exist different form of set membership function and they presented in next figure.

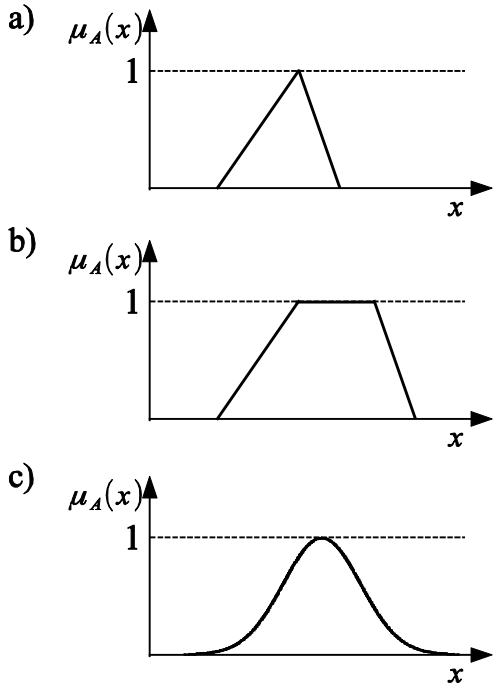

Figure 3: Different form of fuzzy sets: a) triangular, b) trapezoidal, c) exponential

Also, a very important concept in fuzzy logic is a possibility theory [26-27]. That is a complement theory to probability theory to deal with uncertainty. There are two approaches to possibility theory: (i) one, proposed in [26], was to introduce possibility theory as an extension of

fuzzy set theory; (ii) the other, described in [28], was to introduce possibility theory in the framework of Dempster-Schafer's theory of evidence. This approach puts possibility theory in an axiomatic manner.

The major topics in possibility theory are fuzzy arithmetic which is concerned with the operations and computations of fuzzy numbers. Fuzzy numbers are useful for computation in physical sciences and engineering when only imprecise or uncertain sensory data are available for computations. Important notion in the possibility theory is the possibility distribution.

Given fuzzy set A in U and the proposition "x is A" the possibility distribution associated with x, denoted by π_x, is defined to be numerically equal to the membership function of A, that is

$$\pi_x(u) = \mu_A(u) \tag{23}$$

One of the major difference between possibility and probability can be seen from the following

$$P(A) + P(\bar{A}) = 1 \tag{24}$$

$$\Pi(A) + \Pi(\bar{A}) \geq 1 \tag{25}$$

where $P(\cdot)$ is probability and $\Pi(\cdot)$ is possibility.

Possibility theory will be used for design of possibilistic cluster algorithm in the next paragraph.

4. THE ROBUST CLUSTERING AND THE POSSIBILISTIC GUSTAFSON-KESSEL ALGORITHM

Clustering is fundamental part of identification of dynamic systems. Namely, the identification procedure consist of three parts:

(i) Cluster analysis
(ii) Determination of the membership function
(iii) Identification parameters of set of models (off-line or recursive) where number of models equal to number of clusters.

In this paper we consider only of first two items. It is supposed that measurements include outliers which one unavoidable in practice. The clustering is based on optimization and in that sense criterion of clustering must be determined.

Let us denote with μ_{*j} membership of point $\mathbf{x}_j$ in the class of outliers [7]. The membership of point μ_{*j} in the noise cluster is defined to be

$$\mu_{*j} = 1 - \sum_{i=1}^{c} \mu_{ij} \tag{26}$$

where c is the number of clusters and exists condition

$$\sum_{i=1}^{c} \mu_{ij} \leq 1 \tag{27}$$

Criterion for possibilistic cluster method (PCM) is introduced in [9]

$$J_{PC} = \sum_{i=1}^{c}\sum_{j=1}^{N} (\mu_{ij})^m d^2(\mathbf{x}_j, \boldsymbol{\beta}_i) + \sum_{i=1}^{c} r_i \sum_{j=1}^{N} (1-\mu_{ij})^m \tag{28}$$

where N is the number of observations, r_i are positive numbers, $\boldsymbol{\beta}_j$ is the prototype (center of clusters), $m \in (1,\infty)$ (usually $m = 2$) is a weighting exponent and $d(\mathbf{x}_j, \boldsymbol{\beta}_i)$ suitable defined distance. For the Gustafson-Kessel algorithm the distance is Mehalanabis distance

$$d^2(\mathbf{x}_j, \boldsymbol{\beta}_i) = (\det \mathbf{F}_i)^{\frac{1}{n}} (\mathbf{x}_j - \mathbf{c}_i)^T \mathbf{F}_i^{-1} (\mathbf{x}_j - \mathbf{c}_i) \tag{29}$$

where $\mathbf{F}_i$ is the fuzzy covariance matrix of cluster

$$\mathbf{F}_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^m (\mathbf{x}_j - \mathbf{c}_i)(\mathbf{x}_j - \mathbf{c}_i)^T}{\sum_{j=1}^{N} (\mu_{ij})^m} \tag{30}$$

and $\mathbf{c}_i$ centers of clusters ($i = 1,2,\ldots,c$). The algorithm (28)-(30) is the Gustafson-Kessel possibilistic cluster algorithm.

From equation

$$\frac{dJ_{PC}}{d\mu_{ij}} = 0 \tag{31}$$

it follows that

$$\mu_{ij} = \frac{1}{1 + \left[\frac{d^2(\mathbf{x}_j, \boldsymbol{\beta}_i)}{r_i} \right]^{\frac{1}{m-1}}} \tag{32}$$

From (28) and (32) it follows that [7]

$$J_{PC} = \sum_{i=1}^{c}\sum_{j=1}^{N} \left(\frac{1}{1 + \left[\frac{d_{ij}^2}{r_i} \right]^{\frac{1}{m-1}}} \right)^{m-1} \cdot d_{ij}^2 = \sum_{i=1}^{c}\sum_{j=1}^{N} w_{ij} d_{ij}^2 \tag{33}$$

From (33) one can get center of cluster as

$$\mathbf{c}_i = \frac{\sum_{j=1}^{N} w_{ij} \mathbf{x}_j}{\sum_{j=1}^{N} w_{ij}} \tag{34}$$

Using last relation one can to see that the weights in IRLS technique has the role of membership. The quantitative relation between function $\Phi(\cdot)$, $\Psi(\cdot)$ and w and membership function is given in [7]. That is connection between robust statistics and fuzzy logic (relation (34) in [7]). Main differences between algorithms in this paper and reference [7] is in the form of distance d_{ij}.

In stochastic case good alternative for presented methodology are gradient algorithms. When the outliers are present in the measurements and when the membership function has a Gaussian form the gradient algorithm, based on Huber's approach in robust statistics, is presented in

[30]. The case of identification of complex dynamic systems, using Takagi-Sugeno models, is considered in [31].

5. CONCLUSION

In the paper is considered identification of nonlinear dynamic system. The nonlinear system is approximated with fuzzy system (finite collection of Takagi-Sugeno models). The local point in fuzzy identification is determination of membership function. Strategy of determination is based on, owing the outliers presence in measurements, robust clustering. The clustering is based on possibilistic Gustafson-Kessel algorithm. The robustness is based on criterion modification. Also, established the conceptual connection between robust statistic and fuzzy logic.

ACKNOWLEDGEMENTS

This paper is part of projects TR33026 and TR33027 at the University of Kragujevac, Faculty of Mechanical and Civil engineering in Kraljevo, and it was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, Berlin (Germany), (2006)

[2] R. M. Harlick and L.G. Shapiro, "Computer and Robust Vision", Addison-Wesley, New York (USA), (1992)

[3] J. Espinosa, J. Vandewalle and V. Wertz, "Fuzzy logic, identification and predictive control", Springer, Berlin (Germany), (2005)

[4] R. Pearson, "Exploring Data in Engineering, the Science and Medicine", Oxford University Press, Oxford (UK), (2011)

[5] P. Huber and E. Ronchetti, "Robust Statistics", Wiley, New York (USA), (2009)

[6] F. Hampel, P. Rousseenw, E. Ronchetti and W. Stahel, "Robust Statistics. The Approach Based on Influence Functions", Wiley, New York (USA), (1986)

[7] R. Dave and R. Krishnapuram, "Robust Clustering Methods: A Unified View", IEEE Transactions on Fuzzy Systems, Vol. 5, pp. 270-293, (1997)

[8] D. E. Gustafson and W. C. Kessel, "Fuzzy Clustering with a Fuzzy Covariance Matrix", Proceedings of IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes, San Diego (USA), 10-12 January 1979, pp. 761-766, (1979)

[9] R. Krishnapuram and J. Keller, "A Possibilistic Approach to Clustering", IEEE Transactions on Fuzzy Systems, Vol. 1, pp. 98-110, (1993)

[10] D. A. Viattchenin, "A Heuristic Approach to Possibilistic Clustering", Springer, Berlin (Germany), (2013)

[11] F. Lindsten, H. Ohlsson and L. Ljung, "Clustering using Sum-of-norms Regularization with Application to Particle Filter Output Computation", Technical Report: LiTH-ISY-R-2993, Linkoping University, Linkoping (Sweden), (2011)

[12] V. Filipovic and V. Djordjevic, "Fuzzy Cluster Validity Method based on Kolmogorov-Smirnov test", to be published

[13] P. Huber, "Robust Estimation of a Location Parameter", Annals of Mathematical Statistics, Vol. 35, pp. 73-110, (1964)

[14] V. Filipovic, "Recursive Identification of Multivariable ARX Models in the Presence of a Priori Information: Robustness and Regularization", Signal Processing, Vol. 116, pp. 68-77, (2015)

[15] V. Filipovic, "A Global Convergent Outlier Robust Adaptive Predictor for MIMO Hammerstein Models", International Journal of Robust and Nonlinear Control, DOI: 10.1002/rnc.3705, (2016)

[16] V. Filipovic, "Outlier Robust Identification of MIMO ARMAX Models", Asian Journal of Control, Accepted for publication, (2017)

[17] V. Filipovic, "Robust Stochastic Approximation Algorithm for Identification of MIMO Hammerstein OE Models", Nonlinear Dynamics, Accepted for publication, (2017)

[18] B. Grigeliouis, "Student's t-distribution and Related Stochastic Process", Springer, Berlin (Germany), (2013)

[19] L. T. Fernholz, "Von Mises Calculus for Statistical Functionals", Springer, New York (USA), (1983)

[20] F. R. Hampel, "Beyond Location Parameters: Robust Concepts and Methods", Bulletin of the International Statistical Institute, Vol. 46, No. 1, pp. 375-382, (1975)

[21] P. W. Holland and R. E. Welsch, "Robust Regression Using Iteratively Reweighted Least Squares", Communication Statistics – Theory and Methods, Vol. A6, pp. 813-827, (1977)

[22] D. P. O'Leary, "Robust Regression Computation Using Iteratively Reweighted Least Squares", SIAM Journal on Matrix Analysis and Applications, Vol. 11, pp. 466-480, (1990)

[23] K. Chen, Q. Lu, Y. Lu and Y. Dou, "Robust Regularized Extreme Learning Machine for Regression Using Iteratively Reweighted Least Squares", Neurocomputing, Vol. 230, pp. 345-358, (2017)

[24] V. Filipovic, "Robust Recursive Principal Component Analysis in Process Monitoring", to be published, (2017)

[25] L. Zadeh, "Fuzzy set", Information and Control, Vol. 8, pp. 338-353, (1965)

[26] L. Zadeh, "Fuzzy Sets as Basis for a Theory of Possibility", Fuzzy Sets and Systems, Vol. 1, pp. 3-28, (1978)

[27] D. Dubois and H. Prade, "Possibility Theory: An Approach to Computerized Processing of Uncertainty", Plenum Press, New York (USA), (1988)

[28] G. Shafer, “Mathematical Theory of Evidence”, Princeton University Press, Princeton (USA), (1979)

[29] J. Abonyi, “Fuzzy Model Identification for Control”, Birkhauser, Boston (USA), (2003)

[30] V. Filipovic and V. Djordjevic, “Robust Identification of Fuzzy Models Using Gradient Methods”, in preparation, (2017)

[31] V. Filipovic and V. Djordjevic, “Compexity Reduction in Fuzzy System Identification Using Robust Principal Component Analysis”, to be published, (2017)