Empirical comparison of the performance of location estimates of fuzzy number-valued data

Beatriz Sinova[1] and Stefan Van Aelst[2]

[1] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
sinovabeatriz@uniovi.es
[2] Department of Mathematics, KU Leuven, Belgium
stefan.vanaelst@kuleuven.be

Abstract. Several location measures have already been proposed in the literature in order to summarize the central tendency of a random fuzzy number in a robust way. Among them, fuzzy trimmed means and fuzzy M-estimators of location extend two successful approaches from the real-valued settings. The aim of this work is to present an empirical comparison of different location estimators, including both fuzzy trimmed means and fuzzy M-estimators, to study their differences in finite sample behaviour.

Keywords: Fuzzy number · Location · Simulation · Robustness.

1 Introduction

Fuzzy numbers are a useful tool to deal with the imprecision underlying many real-life experiments. For this reason, a methodology to analyze fuzzy number-valued data statistically is of interest and has already provided us with different tools to deal with this kind of data. For example, we could think of regression analysis techniques, clustering, principal components, etc. An important drawback is that a lot of such procedures are based on the use of the Aumann-type mean, which is a generalization of the concept of mean of a random variable. Even when the Aumann-type mean fulfills numerous convenient properties, both from the statistical and probabilistic points of view, it also inherits the lack of robustness of the mean of a random variable. This means that any atypical observation or outlier, or any data changes, may invalidate the conclusions of our study. Unfortunately, it is not uncommon to collect data that include some 'contaminated observations' in real-life experiments. This motivates the search for robust location measures to summarize fuzzy number-valued data sets.

Different robust location alternatives for fuzzy numbers have already been proposed in the literature (see e.g. [1,2,3,4,5]). Among them, the adaptation of trimmed means and M-estimators of location to the fuzzy number-valued settings could be highlighted due to their importance and success for real-valued random variables. The main aim of this paper is to empirically compare the behaviour of fuzzy trimmed means and fuzzy M-estimators in presence of outliers, but other location estimates will also be included in the simulations in order to complete the study.

2 Preliminaries

A (bounded) **fuzzy number** is a mapping $\widetilde{U}:\mathbb{R}\to[0,1]$ such that its α-levels

$$\widetilde{U}_\alpha=\begin{cases}\{x\in\mathbb{R}:\widetilde{U}(x)\geq\alpha\} & \text{if } \alpha\in(0,1]\\ \mathrm{cl}\{x\in\mathbb{R}:\widetilde{U}(x)>0\} & \text{if } \alpha=0,\end{cases}$$

where cl denotes the closure, are nonempty compact convex sets. Therefore, each fuzzy number $\widetilde{U}$ can be uniquely characterized by means of the infima and suprema of all its α-level. $\mathcal{F}_c(\mathbb{R})$ will denote the space of fuzzy numbers.

If $\mathcal{X}$ is a random fuzzy number (a fuzzy number-valued mapping associated with a probability space and such that, for each α, the α-level interval-valued mapping is a random interval associated with the probability space), let $\widetilde{\mathbf{x}}_n=(\widetilde{x}_1,\ldots,\widetilde{x}_n)$ be a sample of fuzzy number-valued observations from $\mathcal{X}$. To represent the central tendency of a data set consisting of several fuzzy numbers, the following measures have been proposed.

- The sample Aumann-type mean [6] is the fuzzy number $\overline{\widetilde{\mathbf{x}}}_n$ such that for all $\alpha\in[0,1]$ its α-levels are given by

$$(\overline{\widetilde{\mathbf{x}}}_n)_\alpha=\left[\sum_{i=1}^n\inf(\widetilde{x}_i)_\alpha/n,\sum_{i=1}^n\sup(\widetilde{x}_i)_\alpha/n\right].$$

- The sample fuzzy trimmed mean [3] is the fuzzy number $\frac{1}{h}\sum_{j\in\widehat{E}_{\widetilde{\mathbf{x}}_\mathbf{n}}}\tilde{x}_j$, where $\widehat{E}_{\widetilde{\mathbf{x}}_\mathbf{n}}$ denotes the corresponding **sample trimming region**, that is,

$$\widehat{E}_{\widetilde{\mathbf{x}}_\mathbf{n}}=\arg\min_{\substack{E\subset\{1,\ldots,n\}\\ \#E=h}}\frac{1}{h}\sum_{i\in E}\left(D_\theta\left(\tilde{x}_i,\frac{1}{h}\sum_{j\in E}\tilde{x}_j\right)\right)^2$$

$$=\arg\min_{E\in\mathcal{E}}Var(\widetilde{\mathbf{x}}_\mathbf{n}|E),$$

with the set $\mathcal{E}=\{E\subset\{1,\ldots,n\}:\#E=h\}$ consisting of all the subsets of h different natural numbers which are up to the sample size, $\theta\in(0,+\infty)$ and D_θ represents the following L^2 metric between fuzzy numbers. Given any $\widetilde{U},\widetilde{V}\in\mathcal{F}_c(\mathbb{R})$,

$$D_\theta(\widetilde{U},\widetilde{V})=\left[\int_{[0,1]}\left(\mathrm{mid}\,\widetilde{U}_\alpha-\mathrm{mid}\,\widetilde{V}_\alpha\right)^2d\ell(\alpha)\right.$$
$$\left.+\theta\int_{[0,1]}\left(\mathrm{spr}\,\widetilde{U}_\alpha-\mathrm{spr}\,\widetilde{V}_\alpha\right)^2d\ell(\alpha)\right]^{1/2},$$

where $\mathrm{mid}\,\widetilde{U}_\alpha=(\inf\widetilde{U}_\alpha+\sup\widetilde{U}_\alpha)/2$ and $\mathrm{spr}\,\widetilde{U}_\alpha=(\sup\widetilde{U}_\alpha-\inf\widetilde{U}_\alpha)/2$.

- The sample M-estimator of location associated with certain loss function ρ [1] is the fuzzy number that minimizes the expression $\frac{1}{n}\sum_{i=1}^{n}\rho(D_\theta(\widetilde{x}_i,\widetilde{U}))$, over $\widetilde{U}\in\mathcal{F}_c(\mathbb{R})$ (if it exists). Concerning the choice of the loss function, Huber's and Hampel's loss functions will be considered along this work. The *Huber loss function*, given by

$$\rho_a^H(x)=\begin{cases}x^2/2 & \text{if } 0\le x\le a\\ a(x-a/2) & \text{otherwise,}\end{cases}$$

with $a>0$ a tuning parameter, is a convex function and puts less emphasis on large errors compared to the squared error loss. On the other hand, the *Hampel loss function* corresponds to

$$\rho_{a,b,c}(x)=\begin{cases}x^2/2 & \text{if } 0\le x<a\\ a(x-a/2) & \text{if } a\le x<b\\ \dfrac{a(x-c)^2}{2(b-c)}+\dfrac{a(b+c-a)}{2} & \text{if } b\le x<c\\ \dfrac{a(b+c-a)}{2} & \text{if } c\le x,\end{cases}$$

where the nonnegative parameters $a<b<c$ allow us to control the degree of suppression of large errors. The smaller their values, the higher this degree. Hampel's family of loss functions is not convex anymore and can better cope with extreme outliers, since observations far from the center ($x\ge c$) all contribute equally to the loss. The following two measures are also particular cases of M-estimators of location.
- The sample 1-norm median [5] is the fuzzy number such that for all $\alpha\in[0,1]$ the corresponding α-level is given by the interval

$$[\mathrm{Me}(\{\inf(\widetilde{x}_i)_\alpha\}_{i=1}^n),\mathrm{Me}(\{\sup(\widetilde{x}_i)_\alpha\}_{i=1}^n)].$$

- The sample wabl/ldev/rdev-median [2] is the fuzzy number such that for all $\alpha\in[0,1]$ the corresponding α-level is given by the interval

$$[\mathrm{Me}(\{\mathrm{wabl}\,\widetilde{x}_i\}_{i=1}^n)-\mathrm{Me}(\{\mathrm{ldev}\,(\widetilde{x}_i)_\alpha\}_{i=1}^n),$$
$$\mathrm{Me}(\{\mathrm{wabl}\,\widetilde{x}_i\}_{i=1}^n)+\mathrm{Me}(\{\mathrm{rdev}\,(\widetilde{x}_i)_\alpha\}_{i=1}^n)],$$

where wabl, ldev_α and rdev_α provide us with an alternative characterization of a fuzzy number. Wabl represents the real number in the interior set $\mathrm{int}(\widetilde{U}_0)$ such that

$$\mathrm{wabl}(\widetilde{U})=\int_{[0,1]}\mathrm{mid}\,\widetilde{U}_\alpha\,d\ell(\alpha)$$

with ℓ the Lebesgue measure, and ldev and rdev functions inform of the left and right deviations w.r.t. wabl, respectively

$$\mathrm{ldev}_{\widetilde{U}}(\alpha)=\mathrm{wabl}(\widetilde{U})-\inf\widetilde{U}_\alpha,$$
$$\mathrm{rdev}_{\widetilde{U}}(\alpha)=\sup\widetilde{U}_\alpha-\mathrm{wabl}(\widetilde{U}).$$

3 Simulation study

This simulation study aims to empirically compare the different alternatives to summarize the central tendency of fuzzy number-valued data in Section 2: fuzzy trimmed means, Huber and Hampel fuzzy M-estimates, 1-norm median and wabl/ldev/rdev-median. In all of them θ is assumed to range in $\{1/3, 1\}$, which are two common choices in the literature (with $\theta = 1/3$ all the points in the α-levels have the same importance for the computation of the D_θ metric, whereas with $\theta = 1$, only the infima and suprema of the α-levels are taken into account). For each of the measures/estimates, the bias, the variance and the mean squared error have been approximated. Different sample sizes ($n = 100, n = 10000$) and different non-contaminated (symmetric and asymmetric) and contaminated distributions have been considered.

Please note that only trapezoidal fuzzy numbers have been considered in order to ease the computation, since a sensitivity analysis has shown that the shape of the fuzzy numbers seems to scarcely affect statistical conclusions (see [7] for more details).

The general scheme of the simulation study is as follows:

Step 1. A sample of n trapezoidal fuzzy number-valued data has been simulated from a random fuzzy number $\mathcal{X}$ for each of the different situations in such a way that

- to generate the trapezoidal fuzzy data, we have considered four real-valued random variables as follows: $\mathcal{X} = \text{Tra}(X_1 - X_2 - X_3, X_1 - X_2, X_1 + X_2, X_1 + X_2 + X_4)$, with $X_1 = \text{mid}\,\mathcal{X}_1$, $X_2 = \text{spr}\,\mathcal{X}_1$, $X_3 = \inf\,\mathcal{X}_1 - \inf\,\mathcal{X}_0$ and $X_4 = \sup\,\mathcal{X}_0 - \sup\,\mathcal{X}_1$ or, alternatively, four ordered real-valued statistics $X_{(1)}, X_{(2)}, X_{(3)}$ and $X_{(4)}$ such that $\mathcal{X} = [X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}]$, i.e., $X_{(1)} = \inf\,\mathcal{X}_0$, $X_{(2)} = \inf\,\mathcal{X}_1$, $X_{(3)} = \sup\,\mathcal{X}_1$ and $X_{(4)} = \sup\,\mathcal{X}_0$;
- each sample is split into a subsample of size $n(1 - c_p)$ (where c_p denotes the proportion of contamination and ranges in $\{0, 0.1, 0.2, 0.4\}$) associated with a non-contaminated distribution and a subsample of size $n \cdot c_p$ associated with a contaminated one, where an additional contamination role is played by C_D (which measures the relative distance between the distribution of the two subsamples and ranges in $\{0, 1, 5, 10, 100\}$);
- 16 situations with different values of c_p and C_D have been considered. For each of these situations two cases have been selected, namely, one in which random variables X_i (or $X_{(i)}$) are independent (CASES 1 and 3) and another one in which they are dependent (CASES 2, 2' and 4).

Step 2. $N = 1000$ replications of *Step 1* have been considered for the situation $c_p = C_D = 0$ in order to approximate the population measures by using a Monte Carlo approach.

Step 3. $N = 1000$ replications of *Step 1* have been considered for all the situations (c_p, C_D) and the approximated estimates, bias, variance and mean squared error have been computed for each location measure.

The choices of the non contaminated and contaminated distributions in each study will be specified now.

Study 1

In the first study, the sample size is $n = 100$, CASE 1 uses

- $X_1 \sim \mathcal{N}(0,1)$ and $X_2, X_3, X_4 \sim \chi^2_1$ for the non-contaminated subsample,
- $X_1 \sim \mathcal{N}(0,3) + C_D$, $X_2, X_3, X_4 \sim \chi^2_4 + C_D$ for the contaminated subsample,

whereas CASE 2 uses

- $X_1 \sim \mathcal{N}(0,1)$ and $X_2, X_3, X_4 \sim 1/(X_1^2+1)^2 + 0.1 \cdot \chi^2_1$ for the non-contaminated subsample,
- $X_1 \sim \mathcal{N}(0,3) + C_D$ and $X_2, X_3, X_4 \sim 1/(X_1^2+1)^2 + 0.1 \cdot \chi^2_1 + C_D$ for the contaminated subsample.

and CASE 2' uses

- $X_1 \sim \mathcal{N}(0,1)$ and $X_2, X_3, X_4 \sim 1/(X_1^2+1)^2 + \sqrt{\chi^2_1}$ for the non-contaminated subsample,
- $X_1 \sim \mathcal{N}(0,3) + C_D$ and $X_2, X_3, X_4 \sim 1/(X_1^2+1)^2 + \sqrt{\chi^2_1} + C_D$ for the contaminated subsample.

Study 2

In the second study, the sample size is $n = 10000$ and in CASES 1, 2 and 2' the distributions for X_1, X_2, X_3 and X_4 in the no-contaminated and the contaminated samples coincide with those for Study 1.

Study 3

In the third study, the sample size is $n = 100$, CASE 3 uses

- $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)} \sim \text{Beta}(5,1)$ (they are simply chosen at random and ordered) for the non-contaminated subsample,
- $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)} \sim \text{Beta}(1, C_D + 1)$ for the contaminated subsample,

whereas CASE 4 uses

- $X_1 \sim \text{Beta}(5,1)$, $X_2 \sim \text{Uniform}[0, \min\{X_1, 1 - X_1\}]$, $X_3 \sim \text{Uniform}[0, X_1 - X_2]$ and $X_4 \sim \text{Uniform}[0, 1 - X_1 - X_2]$ for the non-contaminated subsample,
- $X_1 \sim \text{Beta}(1, C_D + 1)$, $X_2 \sim \min\{X_1, 1 - X_1\} \cdot \text{Beta}(1, C_D + 1)$, $X_3 \sim (X_1 - X_2) \cdot \text{Beta}(1, C_D + 1)$ and $X_4 \sim (1 - X_1 - X_2) \cdot \text{Beta}(1, C_D + 1)$ for the contaminated subsample.

Study 4

In the fourth study, the sample size is $n = 10000$ and in CASES 3 and 4 the distributions for $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}, X_1, X_2, X_3$ and X_4 in the non-contaminated and contaminated samples coincide with those for Study 3.

4 Results

For the bias, variance and mean square error, the conclusions for the different studies are summarized in Tables 1 to 4. The row called Dispersion indicates how variable the choice of the best location measure w.r.t. bias, variance of MSE is ("none" means that the corresponding estimator is the best in all the considered situations; "low", in most of the situations; and "high" if the choice of the best estimator highly depends on the values of the parameters c_p and C_D). For more details about the results, visit http://bellman.ciencias.uniovi.es/SMIRE/Fuzsimul.html.

Table 1. Summary of the main conclusions from Study 1: the best performing (if any) location measures/estimates are indicated for each of the situations

STUDY 1		CASE 1	CASE 2	CASE 2'
Bias	$c_p \leq 0.2$	Hampel	1-norm median	Hampel ($\theta = 1$)
	$c_p = 0.4$	trimmed	1-norm median	trimmed ($\theta = 1$)
	Dispersion	none	low	low
Variance	$c_p = 0$	1-norm median	1-norm median	1-norm median mean
	$c_p \leq 0.2$	1-norm median Hampel ($\theta = 1$)	1-norm median	1-norm median
	$c_p = 0.4$	wabl median 1-norm median	1-norm median	1-norm median
	Dispersion	low	low	high
MSE	$c_p = 0$	Huber ($\theta = 1$) 1-norm median	1-norm median	1-norm median mean
	$c_p \leq 0.2$	Hampel	1-norm median	1-norm median Hampel ($\theta = 1$)
	$c_p = 0.4$	trimmed	1-norm median	trimmed ($\theta = 1$)
	Dispersion	low	low	high

Table 2. Summary of the main conclusions from Study 2: the best performing (if any) location measures/estimates are indicated for each of the situations

STUDY 2		CASE 1	CASE 2	CASE 2'
Bias	$c_p \leq 0.2$	Hampel	1-norm median	Hampel ($\theta = 1$)
	$c_p = 0.4$	trimmed	1-norm median	trimmed ($\theta = 1$)
	Dispersion	none	medium	low
Variance	$c_p = 0$	Huber ($\theta = 1$)	1-norm median	1-norm median
	$c_p \leq 0.2$	Hampel	1-norm median Hampel ($\theta = 1$) trimmed ($\theta = 1/3$)	1-norm median trimmed Hampel ($\theta = 1$)
	$c_p = 0.4$	trimmed	trimmed	trimmed
	Dispersion	medium	low	medium
MSE	$c_p = 0$	Huber ($\theta = 1$)	1-norm median	1-norm median
	$c_p \leq 0.2$	Hampel	1-norm median	Hampel ($\theta = 1$)
	$c_p = 0.4$	trimmed	1-norm median	trimmed ($\theta = 1$)
	Dispersion	none	low	low

5 Conclusions

On the basis of the conclusions gathered in Tables 1, 2, 3 and 4, one can conclude that there is no uniformly most appropriate location estimate. Actually,

Table 3. Summary of the main conclusions from Study 3: the best performing (if any) location measures/estimates are indicated for each of the situations

STUDY 3		CASE 3	CASE 4
Bias	$c_p \leq 0.2$	Hampel ($\theta = 1/3$) trimmed ($\theta = 1/3$)	Hampel ($\theta = 1/3$)
	$c_p = 0.4$	trimmed ($\theta = 1/3$)	trimmed ($\theta = 1/3$)
	Dispersion	none	low
Variance	$c_p = 0$	mean Huber ($\theta = 1/3$)	1-norm median
	$c_p \leq 0.2$	trimmed ($\theta = 1$) wabl median	1-norm median Huber ($\theta = 1/3$)
	$c_p = 0.4$	trimmed ($\theta = 1$)	wabl median trimmed ($\theta = 1/3$)
	Dispersion	high	high
MSE	$c_p = 0$	mean Huber ($\theta = 1/3$)	mean 1-norm median
	$c_p \leq 0.2$	trimmed ($\theta = 1/3$) Hampel ($\theta = 1/3$)	Huber ($\theta = 1/3$) Hampel ($\theta = 1/3$) wabl median
	$c_p = 0.4$	trimmed ($\theta = 1/3$)	trimmed ($\theta = 1/3$) wabl median
	Dispersion	medium	high

Table 4. Summary of the main conclusions from Study 4: the best performing (if any) location measures/estimates are indicated for each of the situations

STUDY 4		CASE 3	CASE 4
Bias	$c_p \leq 0.2$	Hampel ($\theta = 1/3$) trimmed ($\theta = 1/3$)	Hampel ($\theta = 1/3$)
	$c_p = 0.4$	trimmed ($\theta = 1/3$)	trimmed ($\theta = 1/3$)
	Dispersion	none	low
Variance	$c_p = 0$	1-norm median	wabl median
	$c_p \leq 0.2$	trimmed ($\theta = 1$) Hampel ($\theta = 1/3$) 1-norm median	Hampel trimmed ($\theta = 1$) 1-norm median
	$c_p = 0.4$	trimmed ($\theta = 1$)	trimmed ($\theta = 1/3$) 1-norm median
	Dispersion	medium	high
MSE	$c_p = 0$	wabl median	mean wabl median
	$c_p \leq 0.2$	Hampel ($\theta = 1/3$) trimmed ($\theta = 1/3$)	Hampel
	$c_p = 0.4$	trimmed ($\theta = 1/3$)	trimmed ($\theta = 1/3$)
	Dispersion	low	medium

the results seem to indicate that the results depend more on the distributions considered for the non-contaminated and contaminated distributions, or the involved case, than on the sample size. A rather general assertion is that the 1-norm median is the best choice in many cases of Study 1 and Study 2 in terms of any of the considered measures (bias, variance or mean square error), above all in Case 2. In the other cases and studies, the best estimate is not as clear as in Case 2. The Huber and Hampel M-estimators generally behave well for small contamination level while the trimmed means behave well when the proportion of contamination is increased. In Cases 3 and 4, with asymmetric non-contaminated distribution and fuzzy numbers having 0-levels contained in the interval $[0, 1]$, the distinction between the advantages of using these estimates in situations of small or big amounts of contamination is not as evident.

Acknowledgements This research has been partially supported by the Spanish Ministry of Economy, Industry and Competitiveness Grant MTM2015-63971-P. Its support is gratefully acknowledged.

References

1. Sinova, B., Gil, M.Á., Van Aelst, S.: M-estimates of location for the robust central tendency of fuzzy data. IEEE Transactions on Fuzzy Systems **24**(4), 945–956 (2016)
2. Sinova, B.; Pérez-Fernández, S.; Montenegro, M.: The Wabl/Ldev/Rdev Median of a Random Fuzzy Number and Statistical Properties. In: Grzegorzewski, P.; Gagolewski, M.; Hryniewicz, O.; Gil, M.Á. (eds.) Strengthening Links Between Data Analysis and Soft Computing. Advances in Intelligent Systems and Computing, vol. 315, pp. 143–150. Springer (2015)
3. Colubi, A.; González-Rodríguez, G.: Fuzziness in data analysis: Towards accuracy and robustness. Fuzzy Sets and Systems **281**, 260–271 (2015)
4. Sinova, B.; de la Rosa de Sáa, S.; Gil, M.Á.: A generalized L1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data. Information Sciences **242**, 22–34 (2013)
5. Sinova, B.; Gil, M.Á.; Colubi, A.; Van Aelst, S.: The median of a random fuzzy number. The 1-norm distance approach. Fuzzy Sets and Systems **200**, 99–115 (2012)
6. Puri, M.L.; Ralescu, D.A.: Fuzzy random variables. Journal of Mathematical Analysis and Applications **114**, 409–422 (1986)
7. Lubiano, M.A.; Salas, A.; Gil, M.Á.: A hypothesis testing-based discussion on the sensitivity of means of fuzzy data with respect to data shape. Fuzzy Sets and Systems **328**, 54–69 (2017)