Comments on "A distance-based statistical analysis of fuzzy number-valued data" by the SMIRE research group

S Destercke

Comments on "A distance-based statistical analysis of fuzzy number-valued data" by the SMIRE research group

S. Destercke

CNRS, UMR 7253 Heudiasyc, Centre de recherche de Royallieu, 60205 COMPIEGNE

Abstract

This paper is a fine review of various aspects related to the statistical handling of "ontic" random fuzzy sets by the means of appropriate distances. It is quite comprehensive and helpful, as it clarifies the status of fuzzy sets in such methods, explains the advantages of using a distance-based approach, specifies the pitfalls in which one should not fall when dealing with "ontic" random fuzzy sets and provides some illustration of practical computations. Not being a statistician but an occasional user of statistics, my discussion will mainly focus on this more practical aspect.

Keywords. fuzzy data, comments, statistic, ontic view

1. Introduction

First, I must thank Angela Blanco-Fernández, María Rosa Casals, Ana Colubi, Norberto Corral, Marta García-Bárzana, María Ángeles Gil, Gil González-Rodríguez, María Teresa López, María Asunción Lubiano, Manuel Montenegro, Ana Belén Ramos-Guajardo, Sara de la Rosa de Sáa and Beatriz Sinova (henceforth simply denominated "the SMIRE group") for this paper [1], as it has helped me to better understand their approach and ideas.

The SMIRE group has now been working on the issue of statistical analysis of "ontic" fuzzy sets (i.e., fuzzy sets as objects of interest, and not as uncertainty models of a precise yet ill-known quantity) for a long time, and the tools they (among and with

Email address: sebastien.destercke@hds.utc.fr (S. Destercke)

others) develop have now reached a maturity that allows to perform the most standard (and therefore most useful) statistical analysis on fuzzy data.

From a practical standpoint, having such available tools is essential, and the research reported in this paper has allowed to make significant progress in this direction. As a potential user of such tools, there is still some questions whose answer is not entirely clear to me, and that I will therefore discuss here. More precisely, these questions will address the following points:

- What is the advantages of using fuzzy sets as summaries of complex information
- The burden of using a quantitative model and metric
- Issues about interpretation of models and results

2. Fuzzy sets as complex information summary

It is my feeling that, in many case studies handling numerical fuzzy data, such fuzzy data are used as intermediate summaries of complex information (but not always, as for instance in the setting of graded multi-label [2], where fuzzy sets are natural descriptors of the information). For example, they can be used to summarize gradual transitions in an image, rather than using the complete pixel-wise information, or they can summarize the evolution of some variable (e.g., blood pressure [3]) during some period of time. In this sense, they provide more information than summaries consisting of single points (crisp edges, median or mean values, ...), but still provide a compact and perhaps more useful representation of the information than the complete, raw information.

However, as rightly recalled by the SMIRE group in their review, researchers have proposed over the years different ways to summarize such complex information, for instance by using functional data, by the means of histograms or through simple intervals, and have proposed adequate tools to analyze such summarized information. I think that, in most practical cases, one could model the complex information using any of these summaries and apply the corresponding techniques, with more or less difficulty. What I wonder is if there are peculiar types of situations or specific examples where using a fuzzy modeling of the information to summarize it is definitely preferable to the other alternatives, and the reasons that make it preferable?

The case where fuzzy data concerns a subjective source (observer) or a subjective concept (e.g., beauty, quality) is in my opinion quite different, as in this case it is not clear to me to which extent it is legitimate to transform such information into quantitative objects defined on a unique scale, and then draw conclusions from these objects. For instance, in this case the assessment of a same observer over an identical object may change over time, resulting in two different quantitative fuzzy sets, whose "ontic" nature is then not obvious.

3. The burden of numbers

The techniques and approaches reviewed by the SMIRE group heavily rely on the definition of a proper metric (which induces, in some sense, an ordering over fuzzy sets) as well as on the availability of numerical fuzzy sets. While this is fine in a number of situations where obtaining or defining numerical aspects can be done in a satisfying way (I would say most situations involving physical measures, and probably others), I can perceive it as a possible limitation since:

- There are quite a number of statistical tools or notions (the median, rank-based correlation, non-parametric tests such as MannWhitneyWilcoxon) that only use a ranking or an ordering over elements, and that therefore only require a qualitative comparison between fuzzy sets. It is not clear to me if the techniques developed in the paper would apply well to such concepts (I know the SMIRE group has extended the concept of Median to their setting, for instance [4]), and to which extent it would be constraining with respect to a more qualitative view of ordering fuzzy sets;

- In the case of the description of subjective concepts such as beauty or quality, a qualitative notion seems to me more adapted (as suggested by my remark of the previous paragraph): while a numerical translation would, in my opinion, be dangerous to manipulate (since numbers would bear no "ontic" meaning), a more qualitative or ordinal approach would be able to model the information more faithfully (as the same observer will usually be consistent when comparing paintings, i.e., in saying that one painting is more beautiful than another). Note

that, in practice, this could also be obtained by using a numerical inference technique whose conclusions only depend on the ordering of numbers (and not on their particular value), but I do not know if such a technique is available for fuzzy data. Although I agree that fuzzy numbers offer more flexibility (as suggested in Section 3 of the paper), the consequences (in terms of inferences, decision, ...) of manipulating numbers (even if they consist of membership values) as if they had a meaning of their own when they have not are still unclear to me.

It may be the case that the techniques presented in the paper are more adapted to numerical aspects of statistics, in which case building the fuzzy counterpart of qualitative techniques would be another area of research.

4. On the interpretation of inferences and models

One of the main reasons of using statistics is to give a meaning to the data and to communicate it to third-party users. While providing an intuitive interpretation is possible for most techniques of the paper (mean and variance of fuzzy random sets and their comparisons), some other defined statistical values seems to me much harder to interpret. For example:

- The joint covariance $\sigma_{\mathscr{X},\mathscr{Y}}$ is mathematically well-defined and has interesting properties, but its meaning is not clear to me. I understand how to interpret each of its parts $Cov(mid\mathscr{X}_\alpha, mid\mathscr{Y}_\alpha)$ and $Cov(spr\mathscr{X}_\alpha, spr\mathscr{Y}_\alpha)$, as they are covariances of single numbers, but not their combination into $\sigma_{\mathscr{X},\mathscr{Y}}$. How should I interpret a low or a high value of $\sigma_{\mathscr{X},\mathscr{Y}}$? or should I just see it as a convenient mathematical tool without further meaning?

- Similar comments can be done on the linear model $\mathscr{Y} = a\mathscr{X} + \mathscr{E}$, whose parameter estimation itself involves the use of $\sigma_{\mathscr{X},\mathscr{Y}}$. For instance, it is not clear what is the meaning of a "fuzzy noise" $\mathscr{E}$ centered around the fuzzy value B? Can we reduce the model to an expression where the noise is not fuzzy? Is the model equivalent to write $\mathscr{Y} = a\mathscr{X} + B + \mathscr{E}'$ with $\mathscr{E}'$ some noise with zero-mean? Similarly, why considering a fuzzy intercept B but a crisp slope a? Again,

I can understand that such definitions are mathematically attractive and lead to convenient and practical estimation formulas, however they do not necessarily facilitate the interpretation of the model.

It should be noted that the above remark does not mean that such an interpretation does not exist, merely that either more efforts need to be spent in explaining it (in laymen terms), or that the model has to be used as a black-box mapping from inputs to outputs (possibly used to perform predictions).

References

[1] A. Blanco-Fernández, M. R. Casals, A. Colubi, N. Corral, M. García-Bárzana, M. A. Gil, G. González-Rodríguez, M. López, M. A. Lubiano, M. Montenegro, A. B. Ramos-Guajardo, S. de la Rosa de Sáa, and B. Sinova. A distance-based statistical analysis of fuzzy number-valued data. *International Journal of Approximate Reasoning*, This issue, 2014.

[2] W. Cheng, E. Hüllermeier, and K. J. Dembczynski. Graded multilabel classification: The ordinal case. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 223–230, 2010.

[3] M. A. Gil, M. T. López-García, M. A. Lubiano, and M. Montenegro. Regression and correlation analyses of a linear relation between random intervals. *Test*, 10(1):183–201, 2001.

[4] B. Sinova, M. A. Gil, A. Colubi, and S. Van Aelst. The median of a random fuzzy number. the 1-norm distance approach. *Fuzzy Sets and Systems*, 200:99–115, 2012.