A Bayesian modeling approach to fuzzy data analysis

Antonio Calcagnì[1,2] and Przemyslaw Grzegorzewski[3,4]

[1] University of Padova, Italy
[2] GNCS Research Group, National Institute of Advanced Mathematics, Italy
antonio.calcagni@unipd.it

[3] Faculty of Mathematics and Information Science,
Warsaw University of Technology, Poland
[4] Systems Research Institute, Polish Academy of Sciences, Poland
przemyslaw.grzegorzewski@pw.edu.pl

Abstract. Statistical data analysis often entails various uncertainties. Fuzzy numbers help address these complexities, enabling generalized statistical methods. We propose enhancing fuzzy estimators by integrating a general epistemic mechanism. Our approach, validated through simulation studies, offers a flexible solution for fuzzy data analysis.

Keywords: fuzzy statistics, gibbs sampler, fuzzy numbers

1 Introduction

Statistical data analysis frequently involves addressing multiple sources of uncertainty simultaneously. This is particularly evident in the analysis of data from social surveys, where both random components (e.g., sample variation) and systematic components (e.g., such as subjective responses) are intertwined [1]. To disentangle these various sources of uncertainty, fuzzy numbers can be employed, and statistical techniques need to be extended to accommodate the fuzzy representation of the data. Within this framework, specialized methods, like fuzzy Expectation-Maximization (fEM), have been devised to ensure estimation and inference. Nonetheless, due to the construction of epistemic fuzzy estimators, they can suffer from excessive variance [3]. In this contribution, we propose a solution which incorporates the general mechanism assumed to underlie the generation of continuous fuzzy numbers into the definition of fuzzy estimators. The idea relies upon the use of a conditional probabilistic framework that connects the parameters of fuzzy numbers (i.e., the systematic component) to the observed statistical model utilized for data analysis. Consequently, estimation and inference are conducted using the Gibbs sampler-based approach, wherein the full conditional distribution is approximated by sampling from a quadratic approximation of the target posterior distribution [4]. This contribution is structured as follows. Section 2 briefly illustrates the problem we are facing alongside the proposed solution. Section 3 describes the Gibbs sampler derivations whereas

Section 4 illustrates the results of a short simulation study to check the approximation used into the Gibbs schema. Section 5 concludes this short article by providing a summary of its findings.

2 A conditional sampling schema

2.1 Statement of the problem

Let $Y_1, \ldots, Y_n$ be n independent continuous random variables and $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)$ a sample of fuzzy observations. The vector $\tilde{\boldsymbol{y}}$ is a blurred version of $\mathbf{y}$ because of epistemic or *post-sampling* uncertainty-based processes. The interest lies in studying $f_{Y_1,\ldots,Y_n}(\mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{y}})$ with the purpose of making inference on $\boldsymbol{\theta}_{\boldsymbol{y}}$ given the fuzzy sample $\tilde{\boldsymbol{y}}$. From the epistemic perspective on fuzzy statistics, fuzzy observations can be conceptualized as stochastic outcomes influenced by varying degrees of non-random and systematic uncertainty, which obscure the actual, unknown realizations $\mathbf{y}$. In this context, each fuzzy observation $\tilde{y}_i$ consists of mode and precision $\{m_i, s_i\}$ of a Beta-type fuzzy number, which is employed as a general template for representing continuous and unimodal fuzzy numbers [1]. Figure 1 shows two instances of the Beta-type fuzzy number.[1]

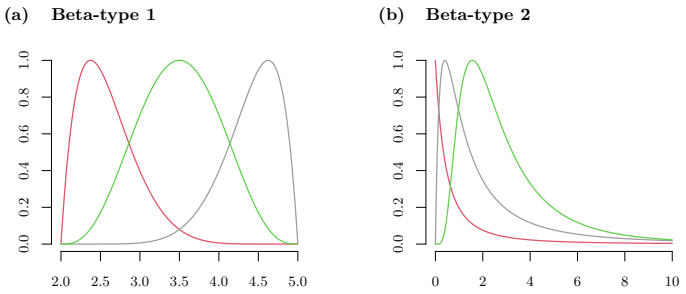

Fig. 1. Some examples of Beta fuzzy sets: (a) bounded case and (b) left-bounded case.

Our objective is to introduce a conditional sampling framework designed to be general enough to handle numerous empirical scenarios involving fuzzy data analysis.

[1] It is important to highlight that Beta-type fuzzy numbers are flexible and parsimonious, requiring only two parameters, while handling with both bounded or semi-bounded variables. Additionally, they serve to generalize triangular fuzzy numbers as well [1].

2.2 Proposed solution

Our main idea revolves around formalizing the two-stage process that governs the generation of (epistemic) fuzzy numbers. Since these stages interact with each other, the concept involves utilizing a conditional probabilistic framework that connects the parameters of fuzzy numbers (such as modes, spreads, and membership function information if available) to the random outcomes of the model $f_Y(y;\boldsymbol{\theta})$. For the general beta representation of fuzzy numbers, the two-stage sampling framework for fuzzy numbers with modes m_i and precision s_i can be expressed as follows

$$y_i \sim f_Y(y;\boldsymbol{\theta_y}), \tag{1}$$

$$s_i \sim \mathcal{G}a(s;\alpha_s,\beta_s), \tag{2}$$

$$m_i|s_i,y_i \sim \begin{cases} \mathcal{B}e_{4P}(m;s_iy_i,s_i-s_iy_1,lb,ub) & \text{if } y_i \in (lb,ub), \\ \mathcal{B}e_P(m;y_i+y_is_i,s_i+2) & \text{if } y_i \in (0,+\infty). \end{cases} \tag{3}$$

In Eq. (1), $f_Y(Y;\boldsymbol{\theta_y})$ depicts the random variable governing the non-fuzzy sampling process, with its parameters expressed as a function of external covariates $\boldsymbol{\theta_y} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$, as in the case of Generalized Linear Models (GLMs). In Eq. (2), the Gamma random variable $\mathcal{G}a$ with parameters $\alpha_s > 0$ and $\beta_s > 0$ is represented, which models the precision (or spread) of the fuzzy number. In the simplest case, s_i is independent of y_i, although this can be generalized to cases where s_i depends on y_i or external covariates as well. Finally, Eq. (3) represents the random variable governing the mode of the fuzzy number, as a function of the true unobserved outcome y_i and the spread s_i. Depending on the context being modeled, it can be defined as a four-parameter Beta distribution ($\mathcal{B}e_{4P}$) or a Beta-prime distribution ($\mathcal{B}e_P$). Note that, in both cases, with the mode-precision parametrization being adopted, the propagation of fuzziness through Eq. (3) results in $(m_1,\ldots,m_n)$ spreading out near $\mathbb{E}[Y]$. Figure 2 illustrates an example of how fuzziness acts on the true unobserved random realization.[2]

3 Inference on $\boldsymbol{\theta_y}$

Given the conditional sampling structure describing the fuzzification process of the outcomes of $Y_1,\ldots,Y_n$, a natural way to make inference about $\boldsymbol{\theta_y}$ involves a kind of *deblurring* procedure, which uses $\tilde{\mathbf{y}}$ instead of the unobserved realizations $\mathbf{y}$. Assuming the observed data is a collection of modes and precisions

[2] Although several choices could have been made for representing the spread (Eq. 2) and mode (Eq. 3) components, the Gamma and Beta distributions exhibit some convenient characteristics in this context. Particularly, Beta distributions, while simplifying the calculus used to obtain the final estimators, also demonstrate great flexibility in modeling unimodal and bounded (or left-bounded) random phenomena. Similarly, the Gamma distribution ensures the positivity of the spread component at a low computational cost.

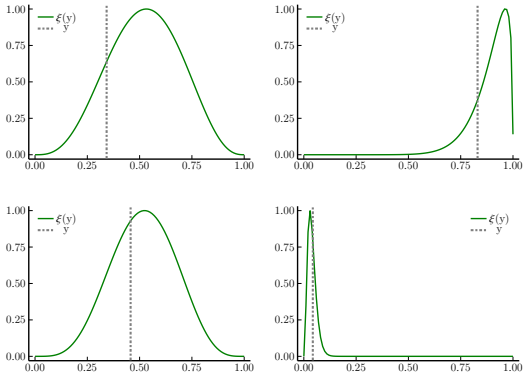

Fig. 2. Examples of the Beta-type 1 fuzzy numbers $\xi_{\tilde{y}}$ (green curves) masking the true unobserved realizations y (dashed vertical gray lines).

$\mathbf{D} = (\{m_1, \ldots, m_n\}, \{s_1, \ldots, s_n\})$, the idea is that of using the Gibbs sampler algorithm. It allows for sampling from the posterior distribution $\pi(\boldsymbol{\theta}_{\boldsymbol{y}}, \mathbf{y}|\mathbf{D})$ by iteratively drawing samples from the conditional posterior distributions ($t > 1$)

$$\mathbf{y}^{(t)} \sim \pi(\mathbf{y}|\mathbf{m}, \mathbf{s}, \boldsymbol{\theta}_y^{(t-1)}), \tag{4}$$

$$\boldsymbol{\theta}_{\boldsymbol{y}}{}^{(t)} \sim \pi(\boldsymbol{\theta}_{\boldsymbol{y}}|\mathbf{m}, \mathbf{s}, \boldsymbol{y}^{(t)}). \tag{5}$$

These are obtained by properly re-arranging the joint distribution as follows

$$\begin{aligned} f(\mathbf{D}, \mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{y}}) &= f(\mathbf{D}|\mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{y}}) f(\mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{y}}) \\ &= f(\mathbf{m}|\mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{y}}) f(\mathbf{s}|\mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{s}}) f(\mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{y}}), \end{aligned}$$

where the prior distribution $f(\boldsymbol{\theta}_{\boldsymbol{y}})$ has been omitted for the sake of simplicity. It is worth noting that due to the independence between Y_i and S_i, the parameters $\boldsymbol{\theta}_{\boldsymbol{s}}$ - in the case of Eq. (3), $\boldsymbol{\theta}_{\boldsymbol{s}} = \{\alpha_s, \beta_s\}$ - can be estimated straightforwardly via maximum likelihood from $\mathbf{s}$. As Eqs. (1)-(3) are quite general and may not lend themselves to analysis within a conjugate Bayesian framework, sampling from the conditional posteriors $\pi(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{y}}, \mathbf{D})$ and $\pi(\boldsymbol{\theta}_{\boldsymbol{y}}|\mathbf{y}, \mathbf{D})$ could pose challenges. Therefore, we opted to adopt a hybrid solution where the step in Eq. (4) is realized via MCMC (e.g., Block sampling algorithm [5]), the step in Eq. (5) is realized via quadratic posterior approximation [4] instead, as described in Section 3.1.

3.1 Approximating $\pi(y|\ldots)$

The quadratic approximation is obtained by matching the parameters of the log posterior density $\pi(y|\ldots)$ with the first K derivatives of a given proposal distribution, which can be chosen according to its feasibility to represent the unnormalized posterior target.

Case 1: $y_i \in (lb, ub)$ In the first case, the proposal posterior density is a four-parameter Beta distribution with mode $\lambda \in (lb, ub)$ and precision $\sigma \in \mathbb{R}^+$. In particular,

$$\begin{aligned}\ln \pi(y_i|\boldsymbol{\theta_y}, \ldots) &\propto \underbrace{-\ln \Gamma(y_i^* s_i) - \ln \Gamma(s_i - s_i y_i^*) + s_i y_i^* \ln\left(\frac{m_i - lb}{ub - m_i}\right)}_{h(y;m,s,lb,ub)} + \ln f_Y(y;\boldsymbol{\theta_y}) \\ &\approx \ln \mathcal{B}e_{4P}(y; \lambda\sigma, \sigma - \sigma\lambda, lb, ub),\end{aligned}$$

where the unknown proposal distribution parameters are found by solving the following equations $(K = 2; k = 1, \ldots, K)$

$$\frac{\partial^k}{\partial y^k} \ln \mathcal{B}e_{4P}(y; \lambda\sigma, \sigma - \sigma\lambda, lb, ub) = \frac{\partial^k}{\partial y^k}\Big(h(y; m, s, lb, ub) + \ln f_Y(y; \boldsymbol{\theta_y})\Big),$$

The solutions $\hat{\lambda}$ and $\hat{\sigma}$ are then used into the first Gibbs step

$$y_i^{(t)} \sim Be_{4P}(y; \hat{\lambda}\hat{\sigma}, \hat{\sigma} - \hat{\sigma}\hat{\lambda}, lb, ub).$$

Case 2: $y_i \in (lb, +\infty)$ In the second case, the proposal posterior density is a Beta prime distribution with mode $\lambda \in (lb, +\infty)$ and precision $\sigma \in \mathbb{R}^+$. In particular,

$$\begin{aligned}\ln \pi(y_i|\boldsymbol{\theta_y}, \ldots) &\propto \underbrace{\ln \mathrm{B}(y_i + s_i, s_i + 2)^{-1} + \ln\left(\frac{m_i}{m_i + 1}\right)(y_i + s_i y_i) + \ln m_i + 2\ln(1 + m_i)}_{g(y;m,s)} \\ &\quad + \ln f_Y(y;\boldsymbol{\theta_y}) \\ &\approx \ln \mathcal{B}e_P(y; \lambda + \lambda\sigma, \sigma + 2),\end{aligned}$$

where the unknown proposal distribution parameters are found by solving the following equations $(K = 2; k = 1, \ldots, K)$

$$\frac{\partial^k}{\partial y^k} \ln Be_P(y; \lambda + \lambda\sigma, \sigma + 2) = \frac{\partial^k}{\partial y^k}\Big(g(y; m, s) + \ln f_Y(y; \boldsymbol{\theta_y})\Big).$$

Similarly to the previous case, the solutions $\hat{\lambda}$ and $\hat{\sigma}$ are used into the first Gibbs step to realize the sampling process

$$y_i^{(t)} \sim Be_P(y; \hat{\lambda} + \hat{\lambda}\hat{\sigma}, \hat{\sigma} + 2).$$

4 Simulation study

In this section, we briefly present the findings of a simulation study conducted to evaluate the quadratic posterior approximation of $\pi(y|\boldsymbol{\theta_y}, \ldots)$ using $\mathcal{B}e_{4P}$ and

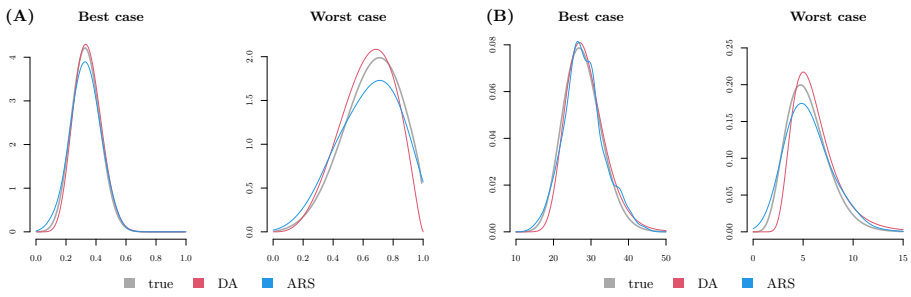

Fig. 3. Simulation study: Reconstructed densities (on average) on Case 1 (panels A) and Case 2 (panels B) in the best and worst result for both DA (red color) and ARS (blue color) algorithms.

$\mathcal{B}e_P$ distributions for both bounded and left-bounded cases.[3] In particular, the derivative-based approximation (DA) has been contrasted against the simple but still effective Adaptive Rejection Sampling (ARS) algorithm for univariate densities [2]. Two measures have been used to assess the process of the approximation, namely the *computation time* (in log scale) and the *total variation distance*.

Two different designs have been used to evaluate the bounded $y \in (lb, ub)$ and left-bounded $y \in (lb, +\infty)$ cases. In particular, in the first case we have used a Logit-Normal distribution for the non-fuzzy model $f_Y(y; \boldsymbol{\theta_y}) = \mathcal{LG}Norm(y; \mu, \phi)$, with $\mu \in \{-1.85, 0, 1.85\}$ and $\phi \in \{1.0, 3.5\}$. The simplest constraint $lb = 0, ub = 1$ has been considered. Instead, in the second case we have used the Gamma distribution as a model of the non-fuzzy component $f_Y(y; \boldsymbol{\theta_y}) = \mathcal{G}a(y; \alpha, \beta)$, with $\alpha \in \{1.0, 1.5, 3.0\}$ and $\beta \in \{4.9.0, 9.0\}$. In both cases, the independent spread component has been modeled as $s \sim \mathcal{G}a(s; 45.0, 45.0/\mu_s)$, with $\mu_s \in \{5.0, 25.0, 50.0\}$ representing high-to-low fuzziness, whereas $n = 2000$ has been set for each combination of simulation factors.

Figures 4-5 show the simulation results across simulation factors for both DA and ARS algorithms. Overall, for the Case 1, the average reconstruction accuracy is 0.97 for DA and 0.95 for ARS, with the average computation log-time being extremely faster for DA (-6.97) then ARS (0.67). Similarly, for the Case 2, the average reconstruction accuracy is 0.93 for DA and 0.85 for ARS, with the average computation log-time being in line with the previous case (DA: -7.27; ARS: 1.15). Finally, Figure 3 illustrates the reconstructed densities in the best and worst result for both algorithms.

5 Conclusions

In this contribution, we proposed a conditional sampling schema for making inference in statistical analyses based on continuous and unimodal fuzzy data.

[3]The overall results concerning the ability of the proposed schema to efficiently sample from the posterior density $\pi(\boldsymbol{\theta_y}, \mathbf{y}|\mathbf{D})$ are left out for the sake of limited space.

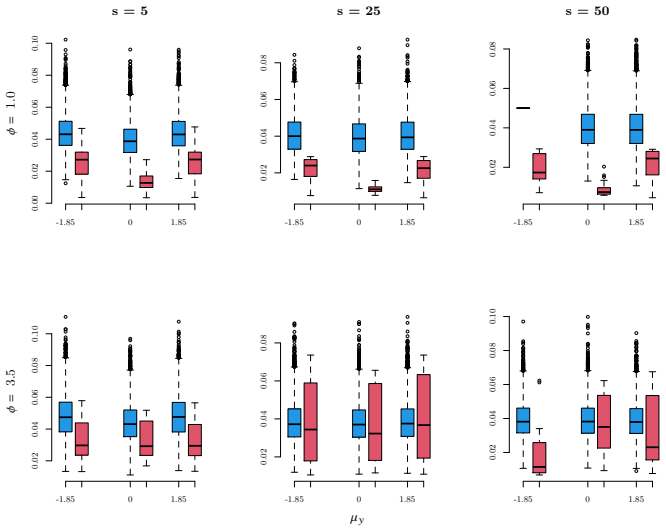

Fig. 4. Simulation study: Case 1 - Total variation distance for each combination of the simulation design for DA (red color) and ARS (blue color) algorithms.

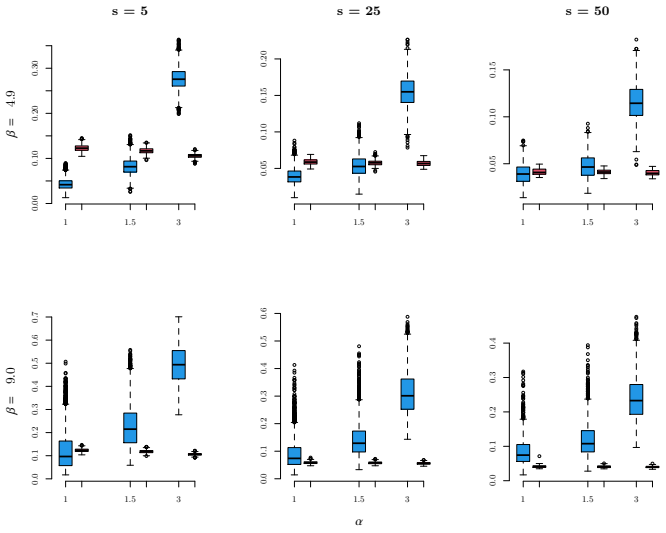

Fig. 5. Simulation study: Case 2 - Total variation distance for each combination of the simulation design for DA (red color) and ARS (blue color) algorithms.

Fuzzy numbers offer a means to address these complexities, but traditional approaches may suffer from high variance. To mitigate this, we propose integrating

a general approach which connects fuzzy parameters to observed statistical models, enabling estimation and inference using the Gibbs sampler-based approach. The proposed solution is based on a hybrid Metropolis within Gibbs schema, where a quadratic-based approximation is also used to mitigate the computation burden of the algorithm. A simulation study has been used to assess the quality of the approximation step in two cases of parameters estimation, one involving the Log-Normal distribution and the other involving the Gamma distributions. The results highlights the effectiveness of the proposed solution if compared against the simple but still efficient Adaptive Rejection Sampling algorithm.

Acknowledgements: Antonio Calcagnì acknowledges the financial support provided through the 2022 Italian MUR grant 2022APAFFN "The attentional curve of forgetting visual information in younger and older adults: experimental and computational insights", which supported this research.

References

1. Calcagnì, A., Cao, N., Rubaltelli, E., Lombardi, L.: A psychometric modeling approach to fuzzy rating data. Fuzzy Sets and Systems **447**, 76–99 (2022)
2. Gilks, W.R., Wild, P.: Adaptive rejection sampling for gibbs sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics) **41**(2), 337–348 (1992)
3. Grzegorzewski, P., Goławska, J.: In search of a precise estimator based on imprecise data. In: 19th World Congress of the International Fuzzy Systems Association (IFSA), 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and 11th International Summer School on Aggregation Operators (AGOP), pp. 530–537. Atlantis Press (2021)
4. Miller, J.W.: Fast and accurate approximation of the full conditional for gamma shape parameters. Journal of Computational and Graphical Statistics **28**(2), 476–480 (2019)
5. Zhou, H., Huang, X.: Bayesian beta regression for bounded responses with unknown supports. Computational Statistics & Data Analysis **167**, 107,345 (2022)