A Bayesian modeling approach to fuzzy data analysis

Antonio Calcagnì
University of Padova

Przemyslaw Grzegorzewski
Warsaw University of Technology

11 International Conference on
Soft Methods in Probability and Statistics
(SMPS24)

September 3, 2024

Introduction

Social surveys are widespread tools used to collect data and insights on people's attitudes, behaviors, and characteristics within a population.

Rating data are a common method for representing information from social surveys, allowing the quantification of opinions and experiences.

Although simple and effective, rating data can introduce uncertainty by not capturing the **complexity of respondents' opinions**.

Introduction

The complexity arises due to the **interplay** of cognitive, affective and contextual factors in the **process of answering** questions using rating scales.

Hence, rating data encapsulate both the rater's **final response** and **epistemic uncertainty**. This type of **post-sampling** uncertainty also coexists with the uncertainty induced by the sampling process.

Fuzzy numbers can be used to mathematically represent this source of uncertainty as **epistemic imprecision** (or **fuzziness**).

Introduction

To deal with fuzziness and randomness appropriately, we need to:

$\Rightarrow$ **Generalize** the statistical modeling to accommodate both sources of uncertainty simultaneously

Introduction

To deal with fuzziness and randomness appropriately, we need to:

⇒ **Generalize** the statistical modeling to accommodate both sources of uncertainty simultaneously

⇒ Consider that **estimators** could suffer from **excessive variance** when **epistemic fuzzy data** are used [Grzegorzewski and Goławska, 2021]

Introduction

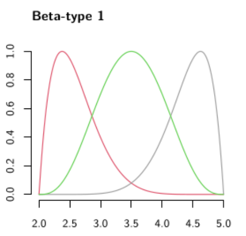
To deal with fuzziness and randomness appropriately, we need to:

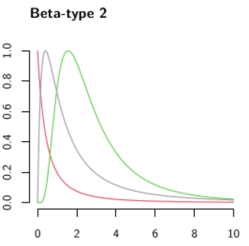
⇒ **Generalize** the statistical modeling to accommodate both sources of uncertainty simultaneously

⇒ Consider that **estimators** could suffer from **excessive variance** when **epistemic fuzzy data** are used [Grzegorzewski and Goławska, 2021]

⇒ Develop a **general** and consistent statistical modeling framework to deal with **fuzzy data analysis**

Introduction





Beta-type fuzzy numbers as a general template for representing **continuous** and **unimodal** fuzzy numbers:

- flexible and parsimonious as they require two parameters only $\{m, s\} \in [lb, ub] \times \mathbb{R}^+$ (mode and precision)
- allow for dealing with variables supported on bounded or semi-infinite intervals (as usual in socio-economic research)
- generalize frequently used fuzzy numbers (triangular, trapezoidal)

Statement of the problem

Let $Y_1, \ldots, Y_n$ be n independent continuous RVs and $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)$ a sample of fuzzy observations. The vector $\tilde{\boldsymbol{y}}$ is a **blurred** realization of $\mathbf{y}$ because of *post-sampling* or *epistemic* uncertainty-based processes.

$\boxed{\text{The interest lies in studying } f_{Y_1,\ldots,Y_n}(\mathbf{y}; \boldsymbol{\theta}_{\boldsymbol{y}}) \text{ with the purpose of making inference on } \boldsymbol{\theta}_{\boldsymbol{y}} \text{ given a fuzzy sample } \tilde{\boldsymbol{y}}.}$

Each fuzzy observation $\tilde{y}_i$ consists of its mode and precision $\{m_i, s_i\}$ of a Beta-type fuzzy number.

A conditional sampling schema

The idea is to use a **conditional schema** linking the statistics of fuzzy numbers to $f_{Y_1,\ldots,Y_n}(\mathbf{y};\boldsymbol{\theta}_\mathbf{y})$:

$$y_i \sim f_Y(y;\boldsymbol{\theta}_\mathbf{y})$$

$$s_i \sim f_S(s;\boldsymbol{\theta}_\mathbf{s})$$

$$m_i|y_i,s_i \sim f_{M|S,Y}(m;\omega(y,s))$$

A conditional sampling schema

$$
\begin{aligned}
\boxed{y_i \sim f_Y(y; \boldsymbol{\theta}_{\boldsymbol{y}})} \\
s_i \sim f_S(s; \boldsymbol{\theta}_{\boldsymbol{s}}) \\
m_i | y_i, s_i \sim f_{M|S,Y}(m; \omega(y, s))
\end{aligned}
$$

RV that governs the stochastic (**non-fuzzy**) sampling process. The parameters can be expressed as a function of external covariates $\boldsymbol{\theta}_{\boldsymbol{y}} = g^{-1}(\mathbf{X}\beta)$ as for GLMs.

The choice of $f_Y(y; \boldsymbol{\theta}_{\boldsymbol{y}})$ depends on the specific problem one is dealing with (e.g., Beta distribution, Logistic distribution, Weibull distribution).

A conditional sampling schema

$$y_i \sim f_Y(y; \boldsymbol{\theta_y})$$
$$\boxed{s_i \sim \mathcal{G}a(s; \alpha_s, \beta_s)}$$
$$m_i | y_i, s_i \sim f_{M|S,Y}(m; \omega(y, s))$$

Gamma distribution with $\alpha_s > 0$ and $\beta_s > 0$ modeling the precision (or spread) of the fuzzy number. In the simplest case, $s_i \perp\!\!\!\perp y_i$ although it can be generalized to cope with cases where s_i depends on y_i or external covariates.

A conditional sampling schema

$$y_i \sim f_Y(y; \boldsymbol{\theta_y})$$
$$s_i \sim \mathcal{G}a(s; \alpha_s, \beta_s)$$
$$\boxed{m_i|y_i, s_i \sim f_{M|S,Y}(m; \omega(y, s))}$$

RV for the mode of the fuzzy number as a function of the true unobserved outcome y_i and the spread s_i.

A conditional sampling schema

$$
\begin{aligned}
y_i &\sim f_Y(y; \boldsymbol{\theta_y}) \\
s_i &\sim \mathcal{G}a(s; \alpha_s, \beta_s) \\
&\boxed{m_i | y_i, s_i \sim f_{M|S,Y}(m; \omega(y,s))}
\end{aligned}
$$

Case 1: $y \in (lb, ub)$, $f_{M|S,Y}(m; \omega(y,s))$ is the **4-parameter Beta distribution**

Case 2: $y \in (0, +\infty)$, $f_{M|S,Y}(m; \omega(y,s))$ is the **Beta prime distribution**

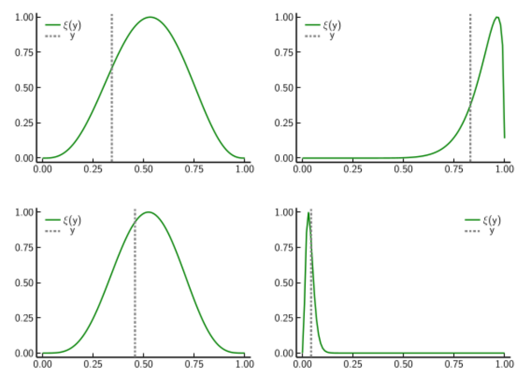
A conditional sampling schema

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

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

$$m_i | s_i, y_i \sim \begin{cases} \mathcal{B}e_{4P}(m; s_i y_i, s_i - s_i y_1, lb, ub), & \text{if } y_i \in (lb, ub) \\ \mathcal{B}e_P(m; y_i + y_i s_i, s_i + 2), & \text{if } y_i \in (0, +\infty) \end{cases} \tag{3}$$

A conditional sampling schema



Examples of a Beta-type 1 fuzzy number $\xi_{\widetilde{y}}$ masking the (true) uncorrupted realizations y

Inference on $\boldsymbol{\theta}_y$

Inference about $\boldsymbol{\theta}_y$ involves a kind of **deblurring** procedure which uses $\tilde{\boldsymbol{y}}$ instead of the unobserved realizations $\mathbf{y}$.

Inference on $\boldsymbol{\theta}_y$

The idea is to plug the hypothesized sampling schema into the estimation procedure, which naturally leads to the **Gibbs sampler**-based solution:

For $t > 1$ do:

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

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

For large T inference on $\boldsymbol{\theta}_{\boldsymbol{y}}$ can be performed by an inspection of the posterior sequence $\left(\boldsymbol{\theta}_{\boldsymbol{y}}^{(1)}, \ldots, \boldsymbol{\theta}_{\boldsymbol{y}}^{(T)}\right)$.

Inference on $\boldsymbol{\theta}_y$

Conditional posterior densities $\pi(\mathbf{y}| \ldots)$ and $\pi(\boldsymbol{\theta}_{\mathbf{y}}| \ldots)$ have unknown forms under the proposed sampling schema. Then, hybrid solutions, such as **posterior approximation** or the **Metropolis within Gibbs** could be used to solve the problem.

Inference on $\boldsymbol{\theta}_y$

<u>Posterior sampling schema</u>

$$\pi(\mathbf{y}, \boldsymbol{\theta}_{\boldsymbol{y}} | \{\mathbf{m}, \mathbf{s}\}) \quad \begin{array}{l} \nearrow \quad \pi(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{y}}, \{\mathbf{m}, \mathbf{s}\}) \quad \text{via a } \textbf{posterior approximation} \\[2em] \searrow \quad \pi(\boldsymbol{\theta}_{\boldsymbol{y}}|\mathbf{y}, \{\mathbf{m}, \mathbf{s}\}) \quad \text{via } \textbf{MCMC} \end{array}$$

Inference on $\boldsymbol{\theta}_y$

Posterior sampling schema

$$\pi(\mathbf{y}|\boldsymbol{\theta}_{\mathbf{y}}, \{\mathbf{m}, \mathbf{s}\}) \quad \text{via a } \textbf{posterior approximation}$$

$$\stackrel{\sim}{=} \mathcal{B}e_{4P}(y; \lambda\sigma, \sigma - \sigma\lambda, lb, ub) \qquad \text{(case 1)}$$

$$\stackrel{\sim}{=} \mathcal{B}e_{P}(y; \lambda + \lambda\sigma, \sigma + 2) \qquad \text{(case 2)}$$

Inference on $\boldsymbol{\theta}_y$

Posterior sampling schema

$\pi(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{y}}, \{\mathbf{m}, \mathbf{s}\})$ via a **posterior approximation**

$$\stackrel{\sim}{=} \mathcal{B}e_{4P}(y; \lambda\sigma, \sigma - \sigma\lambda, lb, ub) \qquad \text{(case 1)}$$

$$\stackrel{\sim}{=} \mathcal{B}e_{P}(y; \lambda + \lambda\sigma, \sigma + 2) \qquad \text{(case 2)}$$

$\{\lambda, \sigma\}$ found by derivative matching [Miller, 2019]

Inference on $\boldsymbol{\theta}_y$

Posterior sampling schema

$\pi(\boldsymbol{\theta}_{\boldsymbol{y}}|\mathbf{y},\{\mathbf{m},\mathbf{s}\})$ via **MCMC**

using the Vihola's Robust Adaptive MH algorithm with a coerced acceptance rate [Vihola, 2012]

Inference on $\boldsymbol{\theta}_y$

Simulation studies show the effectiveness of the approximated hybrid Gibbs sampling to accurately estimate model parameters with a good mixing properties.

Inference on $\boldsymbol{\theta}_y$

Simulation studies show the effectiveness of the approximated hybrid Gibbs sampling to accurately estimating model parameters with a good mixing property.

$\Rightarrow$ Instead, in the next slides we focus on the ability of the proposed conditional schema to reproduce already existing fuzzy data (**external validation**).

Case studies

Aim: Assessing the capability of the proposed conditional schema to reproduce fuzzy data $\tilde{\mathbf{y}}$ collected externally.

Method: Posterior predictive check [Gelman et al., 1996]

i) Fix the non-fuzzy model $f_Y(y; \boldsymbol{\theta})$

ii) Estimate $\hat{\boldsymbol{\theta}}$ using $\tilde{\mathbf{y}}$

iii) Generate B new instances $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_B$ using the conditional schema

Measures: Compare observed statistics $S(\tilde{\mathbf{y}})$ - i.e., *centroids*, 0-*cuts*, *fuzziness* - with the distribution of the simulated ones through the range and Q_3-Q_1 interquartile range.

Case studies

Dataset 1

Dataset: Sample of $n = 69$ observations about Reckless Driving Behavior collected using fuzzy indirect rating scales [Calcagnì and Lombardi, 2022].

Response variable: Driving Anger Scale (DAS) represented as Beta type-1 fuzzy numbers.

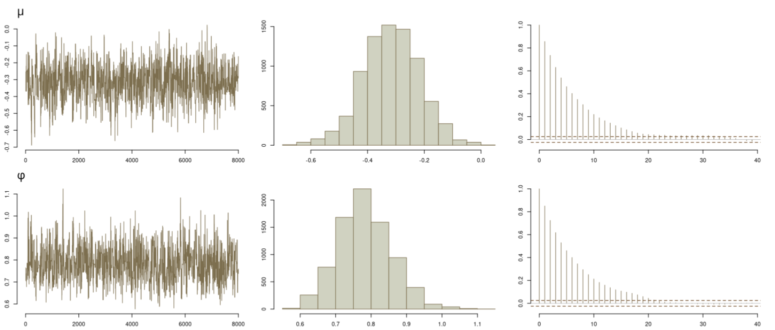
Non-fuzzy model: $f_Y(y;\boldsymbol{\theta}) = \mathcal{L}\text{ogit}\mathcal{N}\text{orm}(y;\mu,\phi)$ with $\{\mu,\sigma\} \in \mathbb{R} \times \mathbb{R}^+$.

Covariates: None.

MCMC: $f(\mu) = \mathcal{N}(;0,100)$, $f(\sigma) = \mathcal{U}(;0,100)$; $\boldsymbol{\theta}_s$ estimated via ML;
No. of samples $1e4$, Burn-in $2.5e3$, acc. rate 0.3045; No. of predictions $B = 5e2$.

Case studies

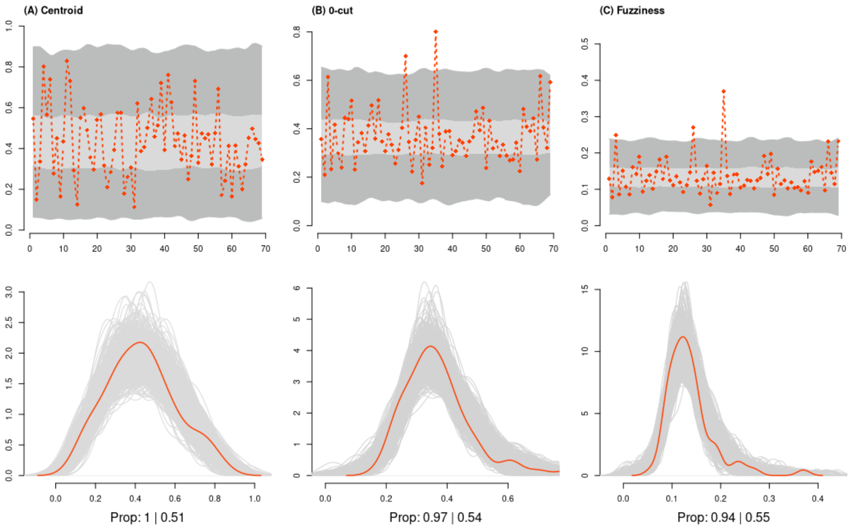
Dataset 1



	Q_1	Median	Mean	Q_3
μ	-0.38	-0.32	-0.32	-0.25
σ	0.73	0.78	0.78	0.83

Case studies

Dataset 1



Case studies

Dataset 2

Dataset: Sample of $n = 49$ observations about Restaurant Quality collected using direct fuzzy rating scales [de Sáa et al., 2014].

Response variable: Restaurant Quality (QR7) represented as Beta type-1 fuzzy numbers (converted from triangular/trapezoidal).

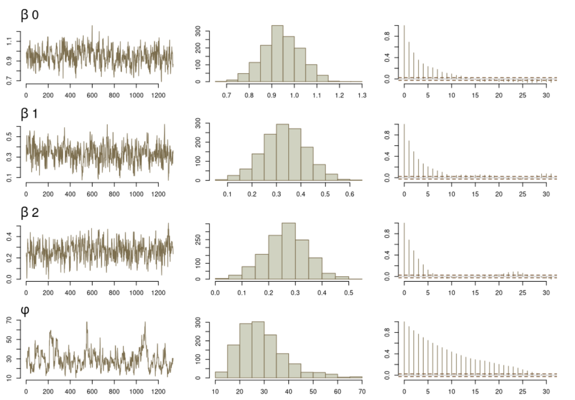
Non-fuzzy model: $f_Y(y;\boldsymbol{\theta}) = Be(y;\boldsymbol{\mu},\phi)$ with $\{\boldsymbol{\mu},\phi\} \in [0,1]^n \times \mathbb{R}^+$
$\mu_i = \text{logit}^{-1}(\mathbf{x}_i\boldsymbol{\beta})$.

Covariates: `quality_food`, `quality_employees` (composite indicators from crisp variables).

MCMC: $f(\beta_j) = \mathcal{N}(;0,100)$ $(j = 1,\ldots,4)$, $f(\phi) = \mathcal{U}(;0,100)$; $\boldsymbol{\theta}_s$ estimated via ML;
No. of samples $1e4$, Burn-in $2.5e3$, acc. rate 0.2554; No. of predictions $B = 5e2$.

Case studies

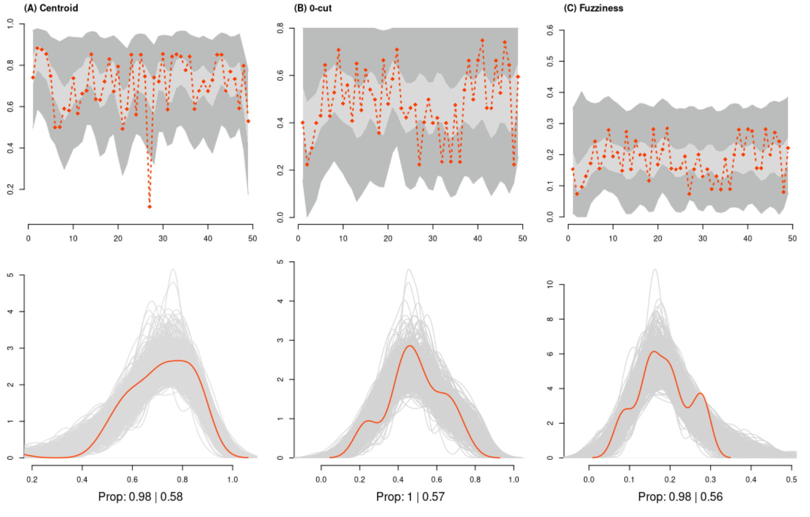
Dataset 2



	Q_1	Median	Mean	Q_3
β_0	0.89	0.94	0.94	1.00
β_1	0.28	0.33	0.33	0.39
β_2	0.21	0.27	0.27	0.32
ϕ	22.36	27.40	29.15	33.92

Case studies

Dataset 2



Case studies

Dataset 3

Dataset: Sample of $n = 147$ observations about Shangai's House Prices collected using fuzzy conversion scales [Zhou et al., 2018].

Response variable: Purchase price represented as Beta type-2 fuzzy numbers (converted from triangular).

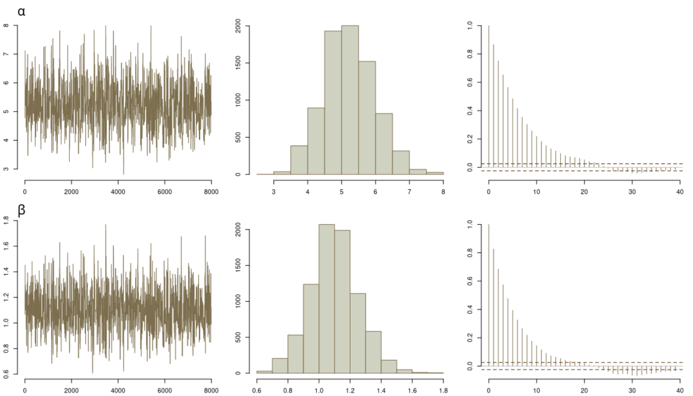
Non-fuzzy model: $f_Y(y;\boldsymbol{\theta}) = Ga(y;\boldsymbol{\alpha},\beta)$ with $\{\alpha,\beta\} \in \mathbb{R}^+ \times \mathbb{R}^+$

Covariates: None.

MCMC: $f(\alpha) = \mathcal{U}(;0,100)$, $f(\beta) = \mathcal{U}(;0,100)$; $\boldsymbol{\theta}_s$ estimated via ML;
No. of samples $1e4$, Burn-in $2.5e3$, acc. rate 0.2876; No. of predictions $B = 5e2$.

Case studies

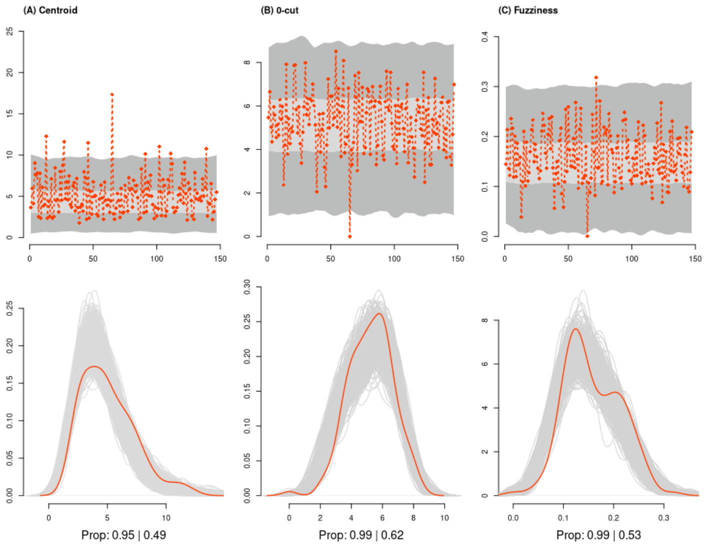
Dataset 3



	Q_1	Median	Mean	Q_3
α	4.70	5.20	5.21	5.69
β	1.00	1.10	1.10	1.19

Case studies

Dataset 3



Concluding Remarks

- Statistical modeling with fuzzy numbers can be of relevant importance in all those situations involving non-stochastic sources of uncertainty (e.g., decision uncertainty in answering a social survey)
- A general and consistent statistical modeling framework to deal with fuzzy data analysis is necessary for practitioners (GLMs-like approach)

Concluding Remarks

- Statistical modeling with fuzzy numbers can be of relevant importance in all those situations involving non-stochastic sources of uncertainty (e.g., decision uncertainty in answering a social survey)
- A general and consistent statistical modeling framework to deal with fuzzy data analysis is necessary for practitioners (GLMs-like approach)
- The proposed schema is entirely probabilistically: fuzziness is summarized into (a few) statistics (e.g., FDA, Network Data Analysis)
- The assumption $s_i \perp\!\!\!\perp y_i$ can be unrealistic in many circumstances

[Calcagnì and Lombardi, 2022] Calcagnì, A. and Lombardi, L. (2022).
Modeling random and non-random decision uncertainty in ratings data: a fuzzy beta model.
AStA Advances in Statistical Analysis, 106(1):145–173.

[de Sáa et al., 2014] de Sáa, S. d. l. R., Gil, M. Á., González-Rodríguez, G., López, M. T., and Lubiano, M. A. (2014).
Fuzzy rating scale-based questionnaires and their statistical analysis.
IEEE Transactions on Fuzzy Systems, 23(1):111–126.

[Gelman et al., 1996] Gelman, A., Meng, X.-L., and Stern, H. (1996).
Posterior predictive assessment of model fitness via realized discrepancies.
Statistica sinica, pages 733–760.

[Grzegorzewski and Goławska, 2021] Grzegorzewski, P. and Goławska, J. (2021).
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)*, pages 530–537. Atlantis Press.

[Miller, 2019] Miller, J. W. (2019).
Fast and accurate approximation of the full conditional for gamma shape parameters.
Journal of Computational and Graphical Statistics, 28(2):476–480.

[Vihola, 2012] Vihola, M. (2012).
Robust adaptive metropolis algorithm with coerced acceptance rate.
Statistics and computing, 22:997–1008.

[Zhou et al., 2018] Zhou, J., Zhang, H., Gu, Y., and Pantelous, A. A. (2018).
Affordable levels of house prices using fuzzy linear regression analysis: the case of shanghai.
Soft Computing, 22:5407–5418.

antonio.calcagni@unipd.it
https://unipd.link/acalcagni