Life is not symmetric: assumptions-free residuals with a BNP approach.

Filippo Ascolani and Valentina Ghidini*

Department of Decision Sciences, Bocconi University

valentina.ghidini@unibocconi.it

1. Introduction ↘

The aim of this work is to exploit a Bayesian Nonparametric framework in order to relax common assumptions for the residuals of models $Y = f(X) + \epsilon$ (namely: homoskedasticity, symmetry and gaussianity), yet retaining the structure and interpretability of common tools.

The advantages of this model are illustrated through an application which aims to predict individual medical costs billed by health insurance in US, based on age, sex, number of children, smoking status, BMI and region of residence.

R package available at: `github.com/ValentinaGhidini/bnpResiduals`

2. BNP Model ←

We start by the standard setting of Bayesian linear models:

$$y_i \mid \beta = X_i'\beta + \epsilon_i, \quad i = 1, \dots, n$$

$$\beta \sim N_p(\cdot \mid b_0, B_0)$$

and we place a BNP prior on the residuals:

$$\epsilon_i \mid \mu_i, \tau_{1i}^2, \tau_{2i}^2 \overset{\text{ind}}{\sim} \frac{1}{2}N(\cdot \mid \mu_i, \tau_{1i}^2) + \frac{1}{2}N(\cdot \mid -\mu_i, \tau_{2i}^2)$$

$$(\mu_i, \tau_{1i}^2, \tau_{2i}^2) \mid P \overset{\text{iid}}{\sim} P$$

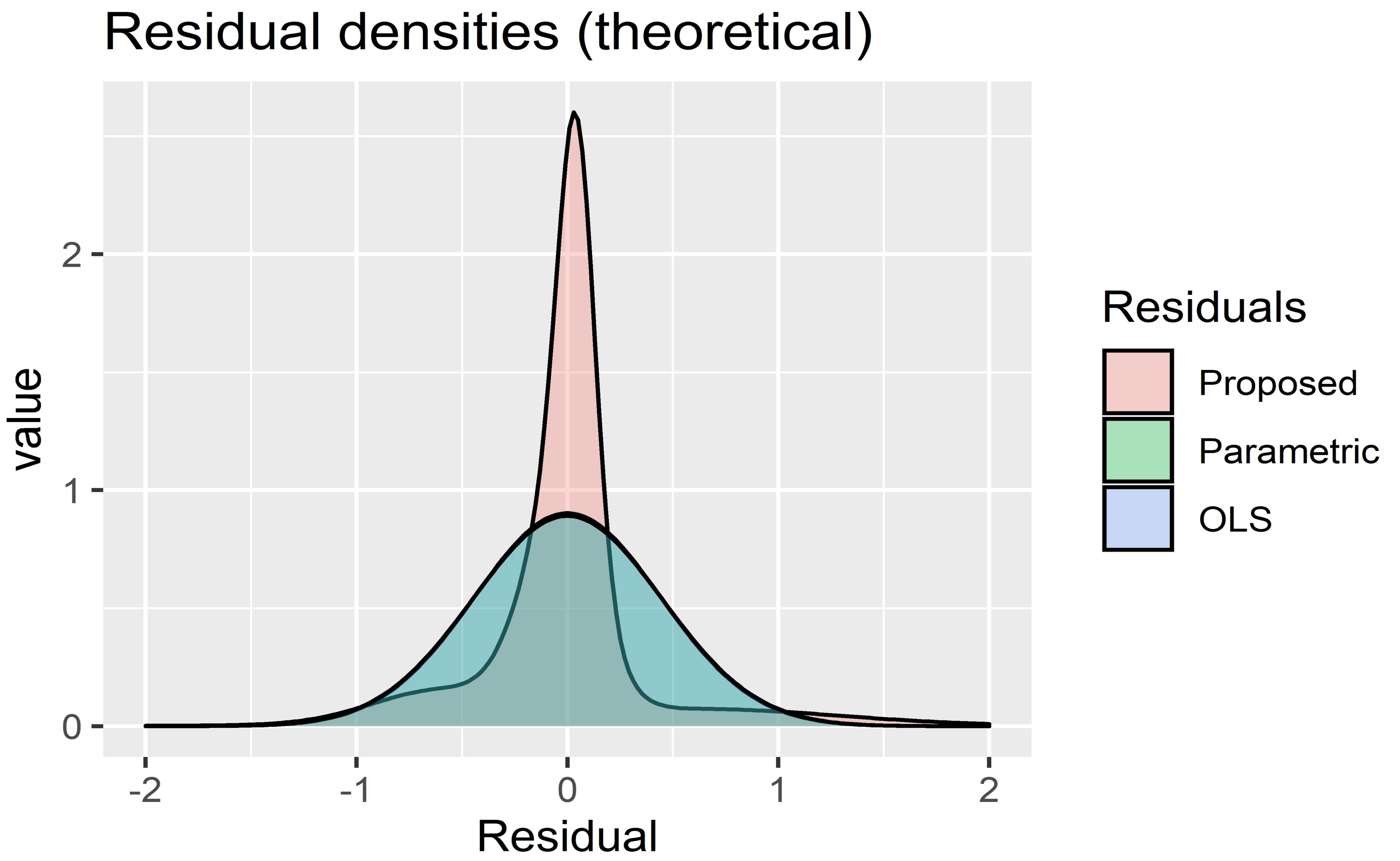
$$\mathrm{P} \sim \mathrm{PY}(P_0, \theta, \sigma)$$

$$\mathrm{P}_0(\cdot) = \mathrm{TN}(\cdot \mid \mu_0, \sigma_0^2) \times \mathrm{IG}(\cdot \mid s_1, S_1) \times \mathrm{IG}(\cdot \mid s_2, S_2),$$

3. Residual Analysis ↓

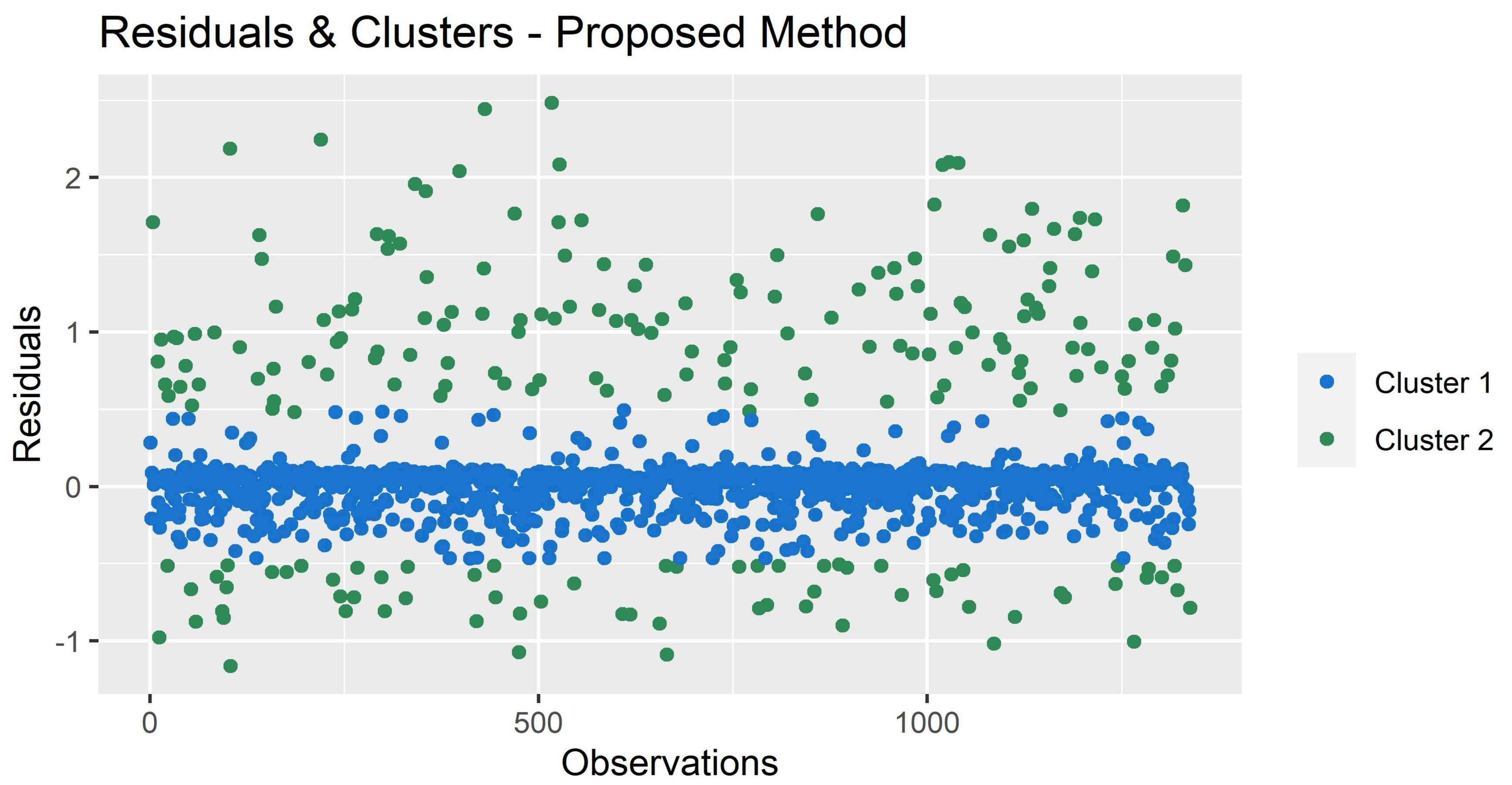
The unconditional density of ϵ_i is given by

$$\epsilon_i \sim \sum_{j \geq 1} W_j \left[\frac{1}{2}N(\cdot \mid \mu_i, \tau_{1i}^2) + \frac{1}{2}N(\cdot \mid -\mu_i, \tau_{2i}^2)\right], \quad (\mu_i, \tau_{1i}^2, \tau_{2i}^2) \overset{\text{iid}}{\sim} P_0.$$



4. Clustering Structure ↓

The model implies that different observations may be sampled from the same component of the countable mixture; in other words, different Y_i may be linked to the same triplet (μ, τ_1, τ_2). In this latent clustering structure, possible outliers can be identified as observations with very different triplets from the others.



5. Robustness - Adding one outlier ↗

Figure 1: Artificial Outlier, added to the southeast residents

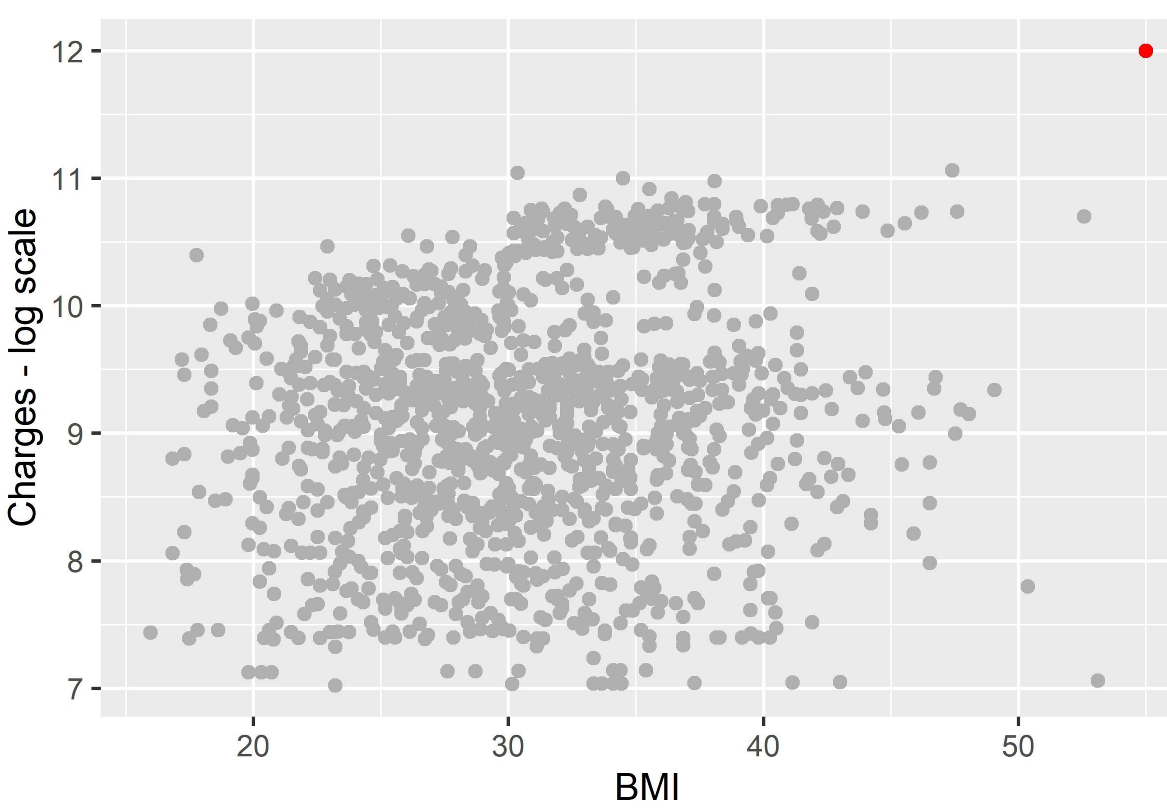


Table 3: Parametric and Nonparametric Coefficients, for the southeast residents (with and without outlier)

	Parametric		Nonparametric	
	without outlier	with outlier	without outlier	with outlier
Sex	0.05	0.18	0.09	0.10
BMI	0.02	0.04	0.00	0.00
Smoker	1.38	1.48	1.69	1.69

The coefficient estimated by the proposed model are much more robust to extreme data.

6. Comparison ↓

Application to US insurance data: hierarchical model for medical costs.

	northeast	northwest	southeast	southwest
Intercept	7.00	6.94	6.73	6.76
Age	0.04	0.04	0.04	0.04
Sex	0.05	0.07	0.09	0.08
BMI	0.00	0.00	0.00	0.00
Children	0.11	0.10	0.11	0.11
Smoker	1.42	1.56	1.69	1.78

Table 1: Nonparametric Coefficients

	northeast	northwest	southeast	southwest
Intercept	6.92	6.98	6.76	6.71
Age	0.03	0.03	0.04	0.04
Sex	0.05	0.08	0.10	0.09
BMI	0.02	0.01	0.01	0.01
Children	0.11	0.12	0.11	0.07
Smoker	1.38	1.41	1.69	1.72

Table 2: Parametric Coefficients

7. Conclusions ↓

- The proposed model relaxes the assumptions of homoskedasticity, symmetry and light tails of the residual distribution.
- It yields an underlying clustering of the residuals, which is useful especially for outliers detection.
- It can be applied to a pletora of models (e.g. ARMA, Gaussian processes etc).
- It provides more robust estimates.

8. References

[1] Ghidini V. Ascolani, F. Bayesian nonparametric residuals. *Forthcoming.*

[2] A. Gelman, J.B Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* 2004.

[3] James L.F. Ishwaran, H. Gibbs sampling methods for stick-breaking priors. *JASA*, 96(453):161–173, 2001.