Robustness under missing data: a comparison with special attention to inference

C. Baum[1*], H. Cevallos-Valdiviezo[2] and A. Van Messem[1]

[1] *Department of Mathematics, University of Liège, Allée de la Découverte 12, 4000 Liège, Belgium; carole.baum@uliege.be, arnout.vanmessem@uliege.be.*
[2] *ESPOL Polytechnic University, Escuela Superior Politécnica del Litoral, ESPOL, Facultad de Ciencias Naturales y Matemáticas, Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador; holgceva@espol.edu.ec.*
[*] *Presenting author*

Keywords. *Robustness; Missing data; Imputation; Inference.*

The size of datasets is increasing at a rapid pace, both in terms of the number of observations as in the amount of included observed characteristics. Along with this, the probability that these datasets contain missing values rises as well. However, certain statistical processes and machine learning techniques are incapable of dealing with incomplete data. As such, it is of the utmost importance that these missing values are dealt with in an adequate way.

Missing value imputation is a highly studied topic. A plethora of techniques have been proposed over the years to find suitable values to replace missing data, ranging from very simple techniques, such as mean or median imputation, to more complicated methods [Lin & Tsai , 2020], such as the popular Multiple Imputation by Chained Equations method (MICE) [van Buuren & Groothuis-Oudshoorn , 2011].

With larger datasets, it is also more likely to observe a number of atypical or extreme data due to measurement and/or encoding errors. These outliers can, to a varying degree, influence statistical analyses. To alleviate this problem, robust techniques have been introduced.

Nowadays, imputation techniques are widely in use, but a large-scale comparison of these methods – and especially in terms of their robustness against outliers – seems to be missing. During a first attempt to fill this gap, we evaluate a large selection of imputation techniques, involving classic and robust procedures, by means of a simulation study with continuous data and different configurations of missing data and outliers. To evaluate the imputation capability and robustness of the imputation techniques we computed the mean prediction error between the actual data values and the predictions obtained by the imputation method. In this study, we also evaluated computational speed of the imputation methods. Our simulations indicate

that, among the single imputation methods, robust linear regression using the MM-estimator and random forest imputation are among the most efficient and robust imputation methods, but these advantages naturally come at a cost, namely a higher computation time.

However, often, the main concern is on the analysis that is performed after imputation. Therefore, in the second phase of our research, we evaluated the inferences and predictions made by different robust regression methods combined with an imputation technique in a simulation study with different configurations of outliers and missing data. For the simulations, we used a similar setting as in Öellerer et al. [2016]. Both rowwise and cellwise outliers were generated, so we considered in the evaluation rowwise robust regression techniques as well as cellwise robust regression techniques. To evaluate the combined regression and imputation strategies in terms of inference capability, we measured the bias and variance of the estimated regression coefficients. To evaluate the prediction capability, we computed the mean prediction error.

References

Stefan van Buuren and Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45**(3), 1–67.

Wei-Chao Lin and Chih-Fong Tsai (2020). Missing value imputation: a review and analysis of the literature (2006-2017). *The Artificial Intelligence Review*, **53**(2), 1487–1509.

Viktoria Öellerer, Andreas Alfons and Christophe Croux (2016). The shooting S-estimator for robust regression. *Computational Statistics*, **31**, 829 - 844.