Stochastic algorithms for robust statistics

Antoine Godichon-Baggioni and Stéphane Robin,
antoine.godichon_baggioni@upmc.fr, stephane.robin@sorbonne-universite.fr
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne-Université, 75005 Paris, France

The acquisition of massive data lying in high dimensional spaces is unfortunately often accompanied by a contamination of these last ones. In this context of contaminated data, even few individuals may corrupt simple statistical indicators such as the mean or the variance. Detecting these atypical data automatically is not straightforward and considering robust techniques is an interesting alternative. There are many robust location indicators Small (1990). For instance, Trimmed-means Rousseeuw and Leroy (2005); Fraiman and Muniz (2001) consist in taking the averaged of the $(1-\alpha)n$ most central information. Nevertheless, this approach necessitates to have an idea of the proportion of contaminated data and assume that these last ones are necessary far from 0. In addition, these approaches often necessitates high computational efforts, although some methods have been developed to deal with dimensionality issues Cuevas et al. (2007).

One should more focus on the geometric median (also called L^1-median or spatial median) introduced by Haldane (1948). Indeed, this location indicator is known to have a 0.5 breakdown point, meaning that even if nearly half of the sample is contaminated, one can control the divergence of the estimates, contrary to the mean which has a 0 breakdown point. In addition, one can also focus on the Median Covariation Matrix, which is the robust alternative to the usual Covariance. In a recent work (Godichon-Baggioni and Robin, 2022), several methods were developed to build robust estimates of the variance. Nevertheless, no theoretical results were given. Then, the objectives of the internship are the following:

1. To give theoretical guarantees for the robust estimates of the covariance obtained with the help of a Weighted Averaged Robbins-Monro procedure (Godichon-Baggioni and Robin, 2022).
2. To derive online robust estimates of the variance (by drawing on Cardot and Godichon-Baggioni (2015)).
3. To obtain theoretical guarantees for these estimates.
4. To apply the proposed methods for online detection of outliers (Rousseeuw and Driessen, 1999).

Lieu du stage: Sorbonne Université, LPSM, 4 place Jussieu, 75005 Paris
Durée: 4 à 6 mois
Gratification: environ 550 euros par mois

References

Cardot, H. and Godichon-Baggioni, A. (2015). Fast estimation of the median covariation matrix with application to online robust principal components analysis. *TEST*, pages 1–20.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *TEST*, 10:419–440.

Godichon-Baggioni, A. and Robin, S. (2022). A robust model-based clustering based on the geometric median and the median covariation matrix. *arXiv preprint arXiv:2211.08131*.

Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*. John wiley & sons.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277.