Can we do without subsampling?

R. Maronna[1*] and V. Yohai[2]

[1] *University of La Plata (rmaronna@retina.ar)*
[2] *University of Buenos Aires and CONICET (victoryohai@gmail.com)*
[*] *Presenting author*

Keywords. *Subsampling; High-dimensional data; Fast estimation; Peña-Yohai regression procedure; Peña-Prieto multivariate estimator*

High breakdown point equivariant estimators in regresssion and multivariate analysis usually require the minimization of a non-convex function of the parameters through an iterative procedure. The possible existence of local minima corresponding to "bad" solutions makes the starting point of the iterations crucial. Subsampling has been the traditional way to compute a robust equivariant starting point. It is known that the number of subsamples required to ensure a given"probabilistic breakdown point" increases exponentially with the data dimension. This fact implies that, except for small dimensionality, there is a conflict between robustness and computing time.

We want to put forward two procedures, one for linear regression and the other for the estimation of multivariate location and scatter, which have been already proposed several years ago for outlier detection, but which have never been considered as initial values for robust iterative estimators. Since their complete description is rather complex, here we describe just the main ideas behind each one.

The first one is a deterministic method proposed by Peña and Yohai (1999) to detect outliers in linear regression. Briefly said, it computes a set of p "sensitivity directions" (where p is the number of parameters) from which it derives $3p$ candidate estimates, which are used as starting values for an S-estimate. The method is very fast. Our simulations show that its use to compute MM-estimates highly improves both robustness and speed.

The second is a semi-deterministic method for multivariate estimation proposed by Peña and Prieto (2007), based on ideas similar to the Stahel-Donoho estimator. But rather than taking purely random directions, they look for two sets directions that have a high probability of exposing outliers. The first set is deterministic, and contains the $2p$ directions that maximize or minimize the kurtosis of the projections.

The other set, which is random, is obtained by an elaborate "stratified sampling", and the number of its elements is proportional to p. These directions are used to compute outlyingness measures for the data points. The complete procedure is rather complex. Simulations by Maronna and Yohai (2015) show that employing this procedure as a starting point highly improves the performance of multivariate MM- and τ-estimators with respect to subsampling, both in robustness and speed.

For these reasons we propose these two procedures to be routinely employed instead of subsampling.

References

Maronna, R. & Yohai, V. (2015). Robust and efficient estimation of high dimensional scatter and location. http://arxiv.org/abs/1504.03389

Peña, D. & Prieto, F.J. (2007). Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, **16**, 228-254.

Peña, D. & Yohai, V. (1999). A Fast Procedure for Outlier Diagnostics in Large Regression Problems. *Journal of the American Statistical Association*, **94**, 434-445.