Considerations on sampling, precision and speed of robust regression estimators

Domenico Perrotta[1], **Marco Riani**[2] and **Francesca Torti**[3]

[1]*European Commission, Joint Research Centre, Ispra, Italy*
[2]*University of Parma, Italy*
[3]*University of Milan Bicocca, Milan, Italy*

Abstract

Methods of very robust regression, which resist up to 50% of outliers, spend a large part of the computational time in sampling subsets of observations and then computing parameter estimates from the subsets. The precision of the estimates depends on the amount of sampling, as we have to find solutions of non-smooth functions with lot of local minima. For example, Least Trimmed Squares (LTS) estimators try to minimize the sum of the h smallest squared residuals, where h is typically $(n-p+1)/2$ and the amount of sampling may vary from one to three thousands of subsets depending on the problem size (see e.g. [2]). To address large datasets, say with $1000 < n < 100.000$ units and $p = 10$ variables, [1] proposed a fast algorithm that can use fewer subsets, but applies c-steps to get approximations with lower objective function value. Moreover, to reduce for large datasets the applications of c-steps, which are $O(n)$, a divide and conquer strategy that partitions the dataset in smaller blocks of 300 observations is used.

We will show how Least Trimmed Squares (LTS) estimators can be made faster with an improved combinatorial sampling approach [3]. Then, we will illustrate the effect of increasing the amount of sampling on the precision of the estimates obtained with the traditional and fast LTS strategies.

Keywords

Least Trimmed Squares, Efficient random samples generation.

References

[1] Rousseeuw, P.J. and K. van Driessen (2006). Computing LTS Regression for Large Data Sets. *Data Min. Knowl. Discov. 12*, 29–45.

[2] Rousseeuw, P.J. and M. Hubert (1997). Recent developments in PROGRESS. *IMS Lecture Notes Monogr. Ser. 31*, 201–214.

[3] Torti, F., D. Perrotta, A.C. Atkinson, and M. Riani (2012). Benchmark testing of algorithms for very robust regression: FS, LMS and LTS. *Comput. Statist. Data Anal.* In press (doi:10.1016/j.csda.2012.02.003).