Quantile regression in large energy datasets

Leo Liberti *CNRS LIX, Ecole Polytechnique,* liberti@lix.polytechnique.fr

A research internship position is announced, for a period of six months at LIX, Ecole Polytechnique, to work on a project called *Quantile Regression for Large Energy Datasets* (QRLED). The QRLED project investigates a new, approximate technique for performing quantile regression computations in extremely large databases. This project received funding by the Siebel Foundation and by Réseau de Transport d'Elécricité (RTE). The Principal Investigator, and internship supervisor, is Leo Liberti. The intern will work in the Data Science and Mining (DASCIM) team at LIX.

Scientific contents

Linear regression is one of the basics of statistics and data science: fitting a hyperplane to a set of n points in $\mathbb{R}^m$ so as to minimize the errors (i.e., distances) between the points and the hyperplane, where "fitting" means "deciding the coefficients": in this sense, classical linear regression is a regression towards the mean. A variant of linear regression can also be carried out with respect to the median and every τ-quantile. Given the m-variate random variables $X_1, \ldots, X_n$, if a linear dependence $a_0 + a_1X_1 + \cdots + a_nX_n = 0$ (†) is postulated between X_n (the dependent variable) and $X_1, \ldots, X_n$ (the independent ones), quantile regression on samples $\tilde{x}_1, \ldots, \tilde{x}_n \in \mathbb{R}^m$ consists in computing the coefficients $a_0, \ldots, a_n$ of (†) such that the τ-quantile of X_n is conditioned on (†). It is well known that linear regression reduces to solving a Linear Program (LP) [1]. We recently discovered that random projections can also be used to approximately solve very large LPs efficiently [2]: our proposal consists in applying this discovery to the quantile regression LP.

What is required of you

The successful candidate will: (a) study and understand the introductory literature electricity prices, quantile regression, and random projections (including the papers cited here); (b) implement efficient codes for generating random matrices and matrix multiplication; (c) write code to read electricity prices data from an external source into memory; (d) use an LP solver API to formulate and solve the corresponding quantile regression LP in memory; (e) use the above code to generate and solve the randomly projected LP; (f) research and test multiple solution retrieval strategies; (g) research and test fast post-processing methods for improving the approximate solution quality. In terms of trade-off between ease of use and efficiency, the best programming language for the tasks above was found to be Julia/JuMP (C/C++ may be even faster, but ease of use will be decreased). Currently, price forecasting in energy transport is a very "hot" applied topic, as evidenced by the two grants supporting this internship. To be successful, you need to dedicate a considerable amount of time and work to this project.

What this internship will give you

If you are successful, you will walk away with enough knowledge to get into most PhD programs in optimization, statistics, data science; and with good chances of being recruited by major energy players worldwide. The internship work will extend for six months in the usual timeframes (March to August or April to September), under the guidance of L. Liberti and C. D'Ambrosio. We are asking for skill and passion in research work (both theoretical and applied). We are legally bound by Ecole Polytechnique to only pay you the minimum internship stipends (around about 530EUR/month). The only means by which we can reward excellence and dedication is by financing study trips abroad (to conferences, seminars and academic visits).

References

[1] R. Koenker. *Quantile regression.* Cambridge University Press, Cambridge, 2005.

[2] K. Vu, P.-L. Poirion, and L. Liberti. Random projections for linear programming. *Mathematics of Operations Research*, accepted.