STOCHASTIC APPROXIMATION TO THE MULTIVARIATE AND THE FUNCTIONAL MEDIAN
Submitted to COMPSTAT 2010

Hervé Cardot[1], Peggy Cénac[1], and Mohamed Chaouch[2]

[1] Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France
herve.cardot@u-bourgogne.fr, peggy.cenac@u-bourgogne.fr
[2] EDF - Recherche et Développement, ICAME-SOAD
1 Av. Général de Gaulle, 92141 Clamart, France
mohamed.chaouch@edf.fr

Abstract. We propose a very simple algorithm in order to estimate the geometric median, also called spatial median, of multivariate (Small, 1990) or functional data (Gervini, 2008) when the sample size is large. A simple and fast iterative approach based on the Robbins-Monro algorithm (Duflo, 1997) as well as its averaged version (Polyak and Juditsky, 1992) are shown to be effective for large samples of high dimension data. They are very fast and only require $O(Nd)$ elementary operations, where N is the sample size and d is the dimension of data. The averaged approach is shown to be more effective and less sensitive to the tuning parameter. The ability of this new estimator to estimate accurately and rapidly (about thirty times faster than the classical estimator) the geometric median is illustrated on a large sample of 18902 electricity consumption curves measured every half an hour during one week.

Keywords: Geometric quantiles, High dimension data, Online estimation algorithm, Robustness, Robbins-Monro, Spatial median, Stochastic gradient averaging

1 Introduction

Estimation of the median of univariate and multivariate data has given rise to many publications in robust statistics, data mining, signal processing and information theory. For instance, the volume of data treated and analyzed by "Electricité De France" (E.D.F.) is getting increasingly important. The installation of systems of measurement becoming more and more efficient, will increase consequently this volume. Our aim will be to have a lighting on these data and information delivered in a current way for a better reactivity about some decision-makings. Then, for example, a rise in competence on the use and modelling of structured data stream should allow the computation

and the analysis of monitoring indicators and performances of the power stations of production in real time, with a data stream environment. Most of the data will be functional data, like load curves for example. Thus, there is a need to have fast and robust algorithms to analyse these functional data. In this context purposes are various, estimation of multivariate central point in a robust way, clustering data around their median, *etc.*

Our work is motivated by the estimation of median profiles with online observations of numerous individual electricity consumption curves which are measured every days at fine time scale for a large sample of electricity meters. The median temporal profile is then a robust indicator of habit of consumption which can be useful for instance for unsupervised classification of the individual electricity demand.

In a multivariate setting different extensions of the median have been proposed in the literature (see for instance Small (1990) and Koenker (2005) for reviews) which lead to different indicators. We focus here on the spatial median, also named geometric median which is probably the most popular one and can be easily defined in a functional framework (Kemperman (1987), Cadre (2001), Gervini (2008)). The median m of a random variable X taking values in some space H ($H = \mathbb{R}^d$, with $d \geq 2$, or a separable Hilbert space) is

$$m \ =: \ \arg\min_{u \in H} \mathbb{E}\left(\|X - u\|\right) \tag{1}$$

where the norm in H, which is the euclidean norm if $H = \mathbb{R}^d$, is denoted by $\|.\|$. The median m is uniquely defined unless the support of the distribution of X is concentrated on a one dimensional subspace of H. Note also that it is translation invariant. The median $m \in H$ defined in (1) is completely characterized by the following gradient equation (Kemperman, 1987),

$$\Phi(m) \ = \ -\mathbb{E}\left(\frac{X - m}{\|X - m\|}\right) = 0. \tag{2}$$

When observing a sample $X_1, X_2, \ldots, X_N$ of N (not necessarily independent) realizations of X, a natural estimator of m is the solution $\widehat{m}$ of the empirical version of (2),

$$\sum_{i=1}^{N} \frac{X_i - \widehat{m}}{\|X_i - \widehat{m}\|} = 0. \tag{3}$$

Algorithms have been proposed to solve this equation (Gower (1974), Vardi and Zhang (2000) or Gervini (2008)). They are needing important computational efforts and can not be adapted directly when data arrive online. For example, the algorithm proposed by Gervini (2008) which is a variant of Gower's approach requires first the computation of the Gram matrix of the data and has a computational cost of $O(N^2 d)$. This also means that a great amount of memory is needed when the sample size N is large. Furthermore

it can not be updated simply if the data arrive online. We propose here an estimation algorithm that can be simply updated and only requires $O(d)$ operations at each step in the multivariate setting. Let us also note that when the data are functional they are generally observed on a common grid of d design points, $\mathbf{X}_i = (X_i(t_1), \ldots, X_i(t_d))$ and then the algorithm will require only $O(d)$ operations at each step, so that it has a global computational cost $O(Nd)$. Let us note that our algorithm is not adapted when one has sparsely and irregularly distributed functional data and this issue deserves further investigation. Note also that in such a large samples context, survey sampling approaches are interesting alternatives (Chaouch and Goga (2010)).

We present in section 2 the stochastic approximation algorithm which is based on the Robbins-Monro procedure. Note that it is very simple and it can be extended directly to the estimation of geometric quantiles (Chaudhuri (1996)). In section 3 a simulation study confirms that this estimation procedure is effective and robust even for moderate sample size (a few thousands). We also remark averaging produces even more efficient estimations. We finally present in section 4 a real study in which we have a sample of $N = 18902$ electricity meters giving every half and hour, during one week, individual electricity consumption and we aim at estimating the temporal median profile.

2 A stochastic algorithm for online estimation of the median

We propose a stochastic iterative estimation procedure which is a Robbins-Monro algorithm (Duflo (1997), Kushner and Yin (2003)) in order to find the minimum of (1). It is based on a stochastic approximation to the gradient of the objective function and leads to the simple iterative procedure

$$\widehat{m}_{n+1} = \widehat{m}_n + \gamma_n \frac{X_{n+1} - \widehat{m}_n}{\|X_{n+1} - \widehat{m}_n\|}, \tag{4}$$

where the sequence of steps γ_n satisfies, $\gamma_n > 0$ for all $n \geq 1$, $\sum_{n\geq 1} \gamma_n = \infty$ and $\sum_{n\geq 1} \gamma_n^2 < \infty$. Classical choices for γ_n are $\gamma_n = g(n+1)^{-\alpha}$, with $0.5 < \alpha \leq 1$. The starting point, m_0 is arbitrarily chosen to be zero.

When α is close to 1, better rates of convergence can be attained at the expense of a larger instability of the procedure so that averaging approaches (Polyak and Juditsky (1992), Kushner and Yin (2003), Dippon and Walk (2006)) have been proposed to get more effective estimators which are less sensitive to the selected values for α and g. When the value of g is a bit too large, averaging also stabilizes the estimator and can reduce significantly its variance. Thus, we also consider an averaged estimator defined as follows

$$\widetilde{m} = \frac{1}{N - n_0} \sum_{n=n_0}^{N} \widehat{m}_n, \tag{5}$$

where n_0 is chosen so that averaging is made on the last ten percent iterations.

Remark 1. Note that this approach can be extended directly to get stochastic approximations to geometric quantiles which are defined as follows by Chaudhuri (1996). Consider a vector $u \in H$, such that $\|u\| < 1$, the geometric quantile of X, say m^u, corresponding to direction u, is defined, uniquely under previous assumptions, by

$$m^u = arg \min_{Q \in H} \mathbb{E} \left(\|X - Q\| + \langle X - Q, u \rangle \right).$$

It is characterized by $\Phi_u(m^u) = \Phi(m^u) - u = 0$, so that one can propose the following stochastic approximation

$$\widehat{m}^u_{n+1} = \widehat{m}^u_n + \gamma_n \left(\frac{X_{n+1} - \widehat{m}^u_n}{\|X_{n+1} - \widehat{m}^u_n\|} + u \right). \tag{6}$$

Remark 2. It can be shown, under classical hypotheses on the distribution of X and the sequence γ_n, that these estimators of the population median and quantiles are consistent. Rates of convergence can also be obtained in the multivariate setting as well as the functional one when H is a separable Hilbert space.

3 A simulation study

We perform simulations in order to check the effectiveness of the algorithm and to evaluate its sensitivity to the tuning parameter g. We have simulated samples of $N = 5000$ brownian motions discretized at $d = 100$ equispaced points in the interval $[0, 1]$. We then added the mean function $m(t) = \sin(2\pi t)$, $t \in [0, 1]$, which is also the median curve for gaussian processes.

Our estimators are defined according (4) and (5) and we take the averaged estimators $\widetilde{m}$ with parameter $n_0 = 500$. They depend on the sequence γ_n. We consider, as it is usually done in stochastic approximation, a sequence defined as follows

$$\gamma_n = \frac{g}{(n+1)^{3/4}}$$

for few different values of $g \in \{0.1, 0.5, 1, 2, 5, 10\}$. The estimation procedure is very fast and computing the geometric median estimator takes less than one second on a PC with the $\mathbb{R}$ language.

We made 100 simulations and evaluate the estimation error with the loss criterion $L(\widehat{m}) = \sqrt{\frac{1}{d} \sum_{j=1}^{d} \left(m(t_j) - \widehat{m}(t_j) \right)^2}$, with $t_j = (j-1)/(d-1)$. We first present in Figure (1) the estimation error for $\widehat{m}_N$ for different values of g and compare it to the error of the empirical mean curve. The iterative estimators are always less effective than the mean curve and their performances depend on the value of the tuning parameter g. In Figure (2) we clearly see that the averaged estimators $\widetilde{m}$ perform really better than the simple ones,

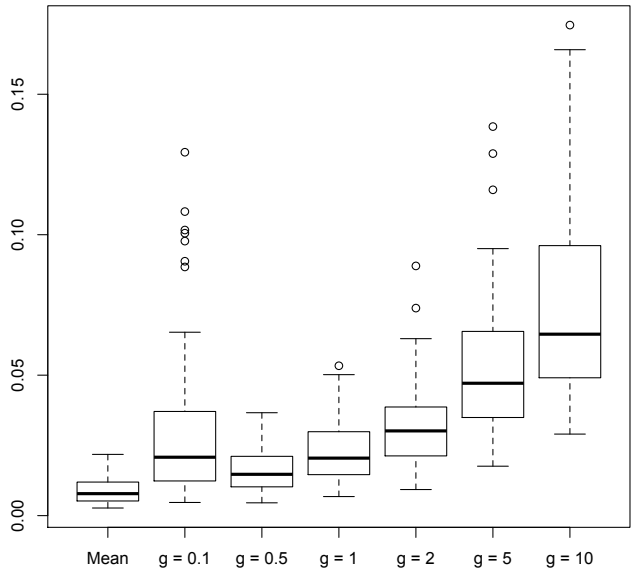

Fig. 1. Approximation error for the mean and the Robbins-Monro estimator of the median for different values of the tuning parameter g, when $\alpha = 0.75$.

with performances which are now comparable to the empirical mean, and do not really depend on g provided that g is not too small.

We also considered the case of a contaminated distribution in which 5% of the observations are also realizations of a brownian with mean function which is now $\mu_c(t) = 5\mu(t)$. The estimation error are presented in Figure (3) and we clearly see that the empirical mean is affected by contamination or outliers whereas the performances of the averaged iterative estimators are still interesting.

As a conclusion of this simulation study, the averaged Robbins-Monro procedure appears to be effective to estimate the geometric median of high dimension data when the sample size is large enough and is not really sensitive to the choice of the tuning parameter g.

4 Estimation of the median electricity consumption curve

We have a sample of $N = 18902$ electricity meters that are able to send electricity consumptions every half an hour during a period of one week,

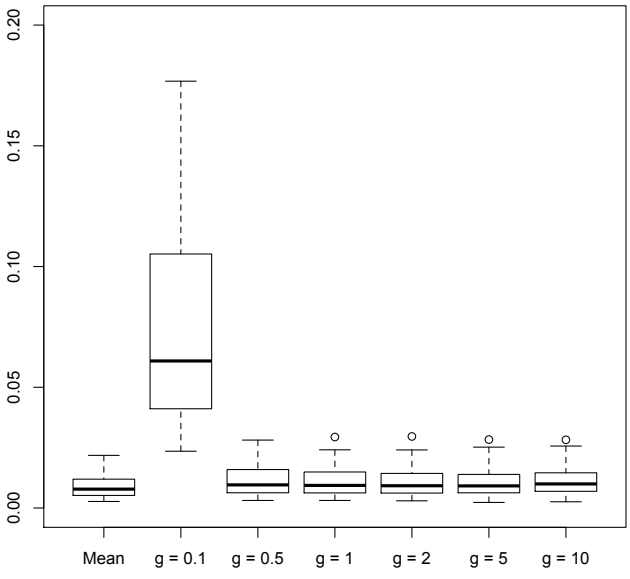

Fig. 2. Approximation error for the mean and the averaged Robbins-Monro estimator of the median for different values of the tuning g when $\alpha = 0.75$.

so that we have $d = 336$ time points. We are interested in estimating the median consumption curve. We present in Figure (4) the estimated geometric median profile for $g = 5$ obtained by averaging the 1000 last iterations and compare it with the mean profile and the pointwise median curve which is obtained by estimating the median value at each instant $t_j, j = 1, \ldots, 336$. The Robbins-Monro estimators are very similar, when averaging, for $g \in [1, 10]$, and different starting points m_0 and are not presented here.

We first remark that there is an important difference between the mean curve and the geometric median curve that is probably due to a small fraction of consumers which have high demands in electricity. There is also a difference, even if it is less important, between the pointwise median and the geometric median and this clearly means that pointwise estimation does not produce the center of our functional distribution according to criterion (1) which takes the following empirical values, 184.3 for the mean function, 173.3 for the pointwise median and 171.7 for the geometric median. The multivariate median was also estimated with the algorithm proposed by Vardi and Zhang (2000) thanks to the function `spatial.median` from the $\mathbb{R}$ package `ICSNP`. The estimated median curve is exactly the same as our but the

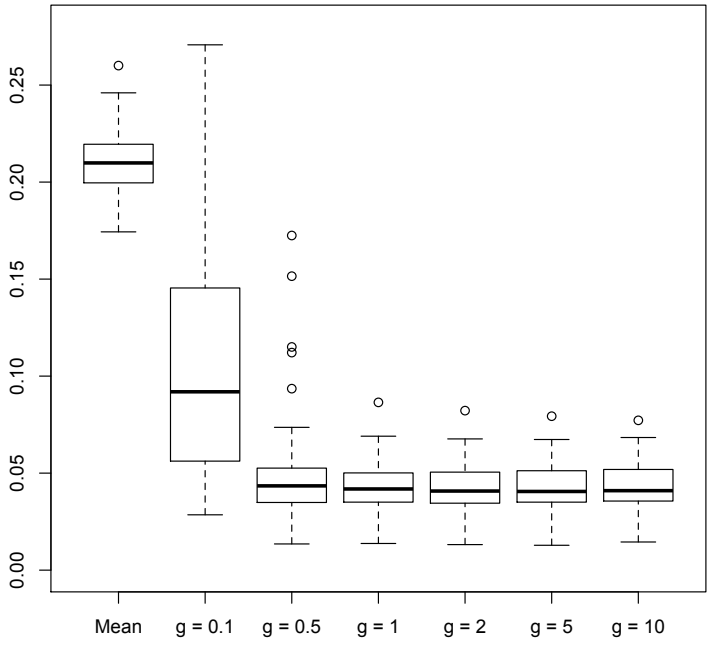

Fig. 3. Estimation error for the mean and the averaged Robbins-Monro estimator of the median for different values of the tuning parameter g when 5% of the data are contaminated.

computation time is much longer (130 seconds versus 3 seconds on the same computer).

References

CADRE, B. (2001). Convergent estimators for the L_1-median of Banach valued random variable. *Statistics*, **35**, 509-521.

CHAOUCH, M., GOGA, C. (2010). Design-Based Estimation for Geometric Quantiles. Accepted for publication in *Comput. Statist. and Data Analysis*.

CHAUDHURI, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, **91**, 862-871.

DIPPON, J., WALK, H. (2006). The averaged Robbins-Monro method for linear problems in a Banach space. *J. Theoret. Probab.* **19**, (2006), 166-189.

DUFLO, M. (1997). *Random Iterative Models*. Springer Verlag, Heidelberg.

GERVINI, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika*, **95,** 587-600.

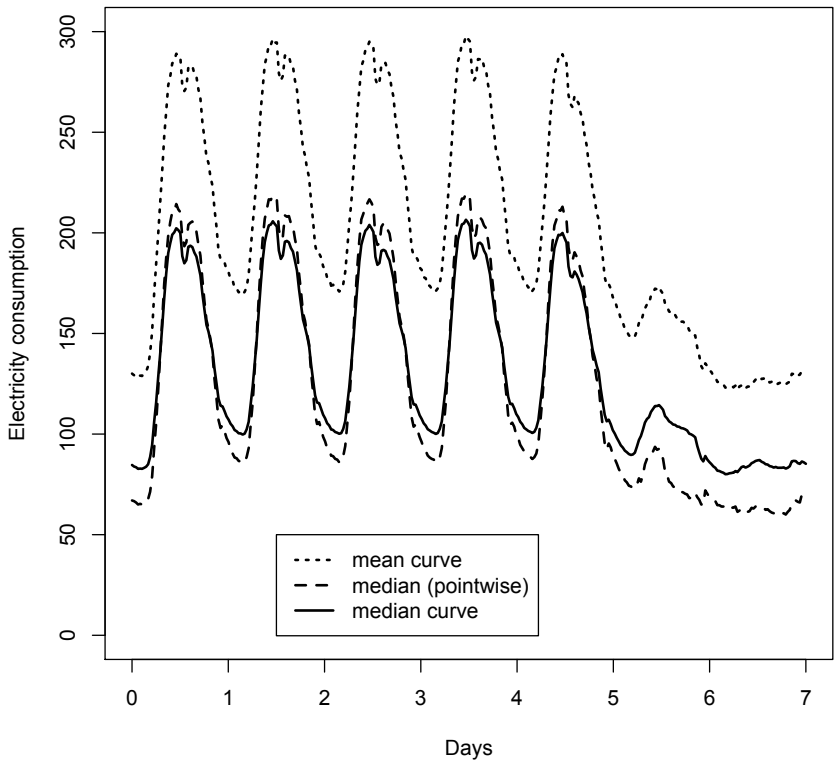

Fig. 4. Comparison of the estimated geometric median profile with the mean electricity consumption curve and the pointwise medians.

GOWER, J.C. (1974). The mediancentre. *Applied Statistics*, **23,** 466-470.

HUBER, P.J., RONCHETTI, E.M. (2009). *Robust Statistics*. John Wiley & Sons, second edition.

KEMPERMAN, J.H.D. (1987). The median of finite measure of a Banach space. In *Statistical data analysis based on the L_1-norm and related methods*, eds Y. Dodge, North-Holland, Amsterdam, 217-230.

KUSHNER, H.J, YIN, G.G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer Verlag, New York.

KOENKER, R. (2005). *Quantile regression*. Cambridge University Press.

POLYAK, B.T., JUDITSKY, A.B. (1992). Acceleration of Stochastic Approximation. *SIAM J. Control and Optimization*, **30**, 838-855.

SMALL, C.G. (1990). A survey of multidimensional medians. *Int. Statist. Inst. Rev.*, **58**, 263-277.

VARDI, Y., ZHANG, C.H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, **97**, 1423-1426.