Fast and efficient estimation of the geometric median in high dimension : a stochastic gradient approach

Hervé Cardot

Institut de Mathématiques de Bourgogne, Université de Bourgogne
with Peggy Cénac (Univ. Bourgogne) and Mohamed Chaouch (EDF)

Herve.Cardot@u-bourgogne.fr

Compstat - August 2010 - Paris

The median in $\mathbb{R}$

A "central" notion in statistics since Laplace.

For a random variable taking values in $\mathbb{R}$:

"the" (may not be unique) value m such that $\mathbb{P}(X \leq m) = 0.5$.

Another characterization of the median m

$$\mathbb{E}\left(\text{sign}(X - m)\right) \;=\; \int \text{sign}(X(\omega) - m) dP(\omega) \;=\; 0.$$

Since $\text{sign}(X - m) = \frac{X-m}{|X-m|}$, we also have the following characterization,

$$m \;=\; \arg\min_{z \in \mathbb{R}} \; \mathbb{E}\,|X - z| \;.$$

- The quantile of order α, for $\alpha \in]0, 1[$, is defined by $\mathbb{P}(X \leq q_\alpha) = \alpha$.

Equivalently,

$$q_\alpha \;=\; \arg\min_{z \in \mathbb{R}} \; \mathbb{E}\left[|X - z| + (2\alpha - 1)(X - z)\right].$$

The geometric median in $\mathbb{R}^d$ (or in a separable Hilbert space H)

In the Euclidean space $\mathbb{R}^d$ equipped with its usual norm $\| \ \|$, a natural generalization of the median

$$m \ := \ \arg\min_{z \in H} \ \mathbb{E}\|X - z\|,$$

called geometric or spatial median (Haldane, 1948).

Property (Kemperman, 1987)
If the space H is strictly convex, the geometric median m is unique, unless the support of X is within a one dimensional subspace.

- Examples of strictly convex spaces :
- Euclidean spaces $\mathbb{R}^d$, when $d > 1$,
- separable Hilbert spaces H,
- some Banach spaces (L_p, $1 < p < \infty$).

- Support condition :
$\exists (u_1, u_2) \in H \times H$, $\langle u_1, u_2 \rangle = 0$, $\mathbb{V}\text{ar}(\langle u_1, X \rangle) > 0$ and $\mathbb{V}\text{ar}(\langle u_2, X \rangle) > 0$.

Characterization of the geometric median

We suppose there are no atoms ($\forall x \in H,\ \mathbb{P}(X = x) = 0$).

Then $G : H \mapsto \mathbb{R}$ defined by $G(x) = \mathbb{E}\|X - x\|$, is strictly convex and Fréchet différentiable

$$\Phi(x) \ := \ \nabla G_x \ = \ -\mathbb{E}\left(\frac{X - x}{\|X - x\|}\right).$$

The median m is characterized by $\nabla G_m = 0$.

G has a second order Fréchet derivative, at point m, $\Gamma_m : H \mapsto H$,

$$\Gamma_m \ := \ \mathbb{E}\left[\frac{1}{\|X - m\|}\left(I_H - \frac{(X - m) \otimes (X - m)}{\|X - m\|^2}\right)\right],$$

where I_H is the identity operator and $u \otimes v = \langle u, . \rangle v$, for $(u, v) \in H^2$.

If $\mathbb{E}\|X - m\|^{-1} < \infty$, then Γ_m is a bounded and strictly positive operator. There are constants, $\infty > \mathbb{E}\|X - m\|^{-1} = \lambda_M > \lambda_m > 0$,

$$\lambda_M \|u\|^2 \geq \langle \Gamma_m u, u \rangle \geq \lambda_m \|u\|^2, \quad \forall u \in H.$$

Robustness : the influence function

Consider a distribution P_0 contaminated by a point-mass distribution at $z \in H$,

$$P_{\epsilon,z} = (1-\epsilon)P_0 + \epsilon\delta_z.$$

The influence function

$$\mathrm{IF}_m(z) = \lim_{\epsilon \longrightarrow 0} \frac{m(P_{\epsilon,z}) - m(P_0)}{\epsilon}$$

is a measure of the sensitivity of the median m to a small perturbation of the target distribution.

Property

$$IF_m(z) = \Gamma_m^{-1} \frac{z-m}{\|z-m\|}$$

and the gross error sensitivity is bounded as follows

$$\sup\{\|IF_m(z)\|,\ z \in H\} = \frac{1}{\lambda_m}.$$

- The gross error sensitivity is not bounded for the mean.

Estimation in $\mathbb{R}^d$

Suppose we have a sample of n independent realizations, $X_1, \ldots, X_n$. The usual estimator of m (Gower, 1974, Vardi & Zhang, 2000, Gervini, 2008) is characterized by

$$\sum_{i=1}^{n} \frac{X_i - \widehat{m}}{\|X_i - \widehat{m}\|} = 0,$$

Iterative and rather long procedures are needed (Newton-Raphson or Weiszfeld) to get a numerical solution

$$\sum_{i=1}^{n} \frac{X_i - \widehat{m}}{\|X_i - \widehat{m}\|} = 0 \Rightarrow \widehat{m}^{e+1} = \sum_{i=1}^{n} p_i(\widehat{m}^e)\, X_i.$$

Property (Haberman, 1989, Niemiro, 1992).
If $H = \mathbb{R}^d$, when $n \to +\infty$,

$$\sqrt{n}\,(\widehat{m}_n - m) \rightsquigarrow \mathcal{N}(0, \Gamma_m^{-1} Var(S(X - m))\Gamma_m^{-1})$$

with $S(u) = u/\|u\|$, $u \in \mathbb{R}^d$.

A very simple and very fast algorithm

We consider the algorithm

$$m_{n+1} = m_n + \gamma_n \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}$$

and suppose the steps γ_n are such that $\forall n, \gamma_n > 0$,

$$\sum_{n \geq 1} \gamma_n = \infty \quad \text{and} \quad \sum_{n \geq 1} \gamma_n^2 < \infty.$$

Advantages

- For a sample of n realizations in $\mathbb{R}^d$: $O(nd)$ operations.
- No need to store all the data (data streams)
- Automatic update (online estimation)

A Robbins-Monro (1951) algorithm

This algorithm (stochastic gradient) can also be written

$$m_{n+1} = m_n - \gamma_n \big(\underbrace{\Phi(m_n)}_{\text{gradient}} + \zeta_{n+1} \big),$$

with $\zeta_{n+1} = -\frac{X_{n+1}-m_n}{\|X_{n+1}-m_n\|} - \Phi(m_n)$.

- If the X_n are $i.i.d.$, the sequence ζ_{n+1} is a martingale difference,

$$\mathbb{E}\left(\zeta_{n+1} \mid \mathcal{F}_n\right) = 0 \quad \text{avec } \mathcal{F}_n = \sigma(X_0, \ldots, X_n).$$

Moreover,

$$\mathbb{E}\left(\|\zeta_{n+1}\|^2 | \mathcal{F}_n\right) \leq 4.$$

A remark on geometric quantiles estimation

This approach can be extended directly to get stochastic approximations to geometric quantiles (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

$$\Phi_u(m^u) = \Phi(m^u) - u = 0.$$

The following stochastic approximation

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

A convergence result in Hilbert spaces

Result (Cardot, Cénac, Zitt 2010)
The sequence m_n converges almost surely when n tends to infinity,

$$\|m_n - m\| \to 0, \ \ p.s.$$

- Sketch of the proof (classical)

Show that for all $\epsilon \in]0,1[$, $\mathbb{P}(\Omega_\epsilon) = 0$, with

$$\Omega_\epsilon = \left\{\omega \in \Omega \quad : \ \exists n_\epsilon(\omega) \geq 1, \ \forall n \geq n_\epsilon(\omega), \quad \epsilon < V_n(\omega) < \epsilon^{-1}\right\}$$

considering that

$$\lim_{n\to\infty} \mathbb{E}V_{n+1} = \mathbb{E}V_0 + \lim_{n\to\infty} \left(\sum_{j=1}^{n} \gamma_j^2 + 2\sum_{j=1}^{n} \gamma_n \ \mathbb{E}\underbrace{\langle \Phi(m_n), m - m_n\rangle}_{<-\lambda_\epsilon \epsilon \text{ in } \Omega_\epsilon} \right) \leq C$$

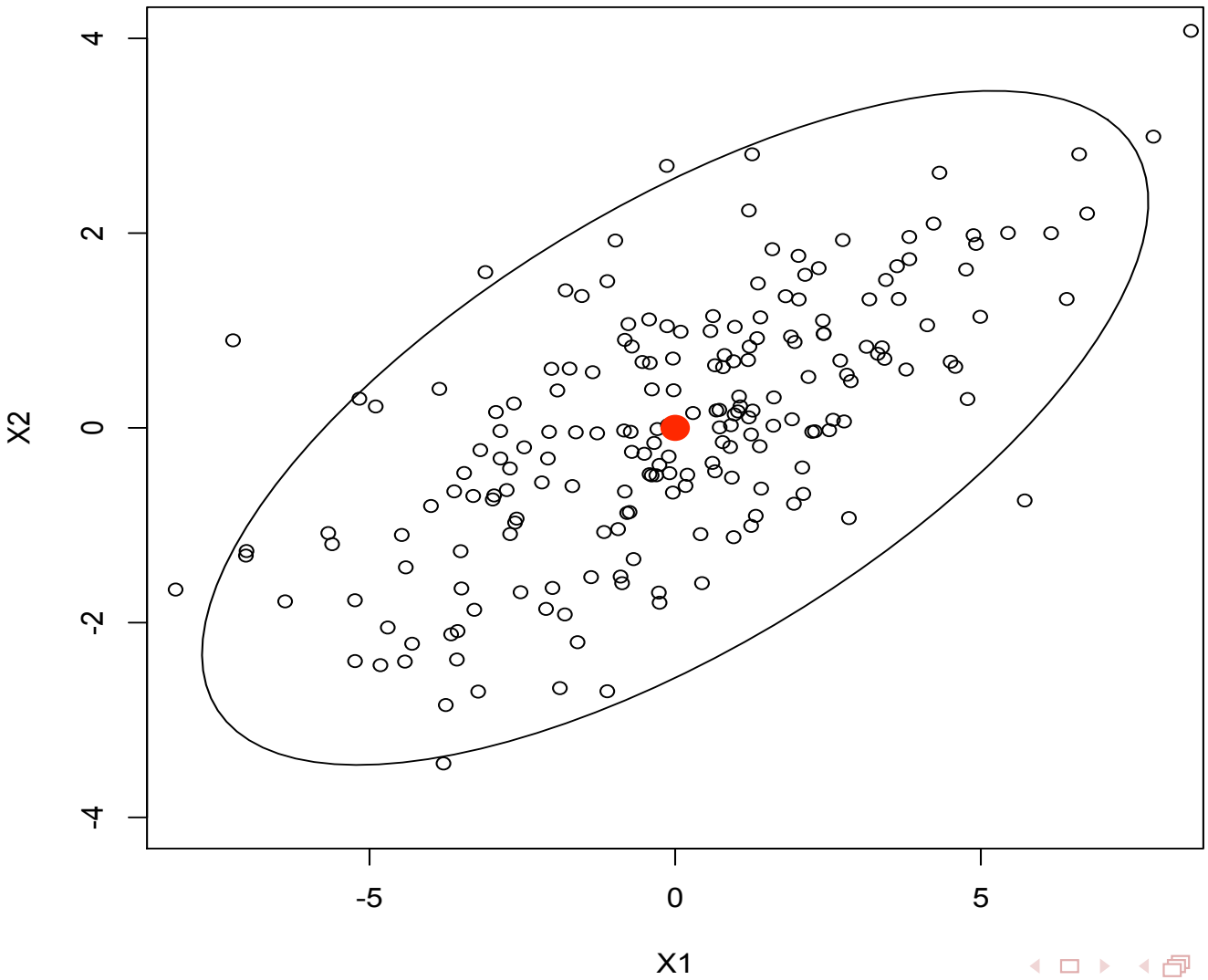
Property
For all $r > 0$, the function G restricted to $x \in B(m,r)$ is strongly convex : $\exists \lambda_r > 0$, such that $\forall x \in B(m,r)$

$$\langle \Phi(x), x - m\rangle \geq G(x) - G(m) \geq \lambda_r \|x - m\|^2 \ .$$

Does it really work ?

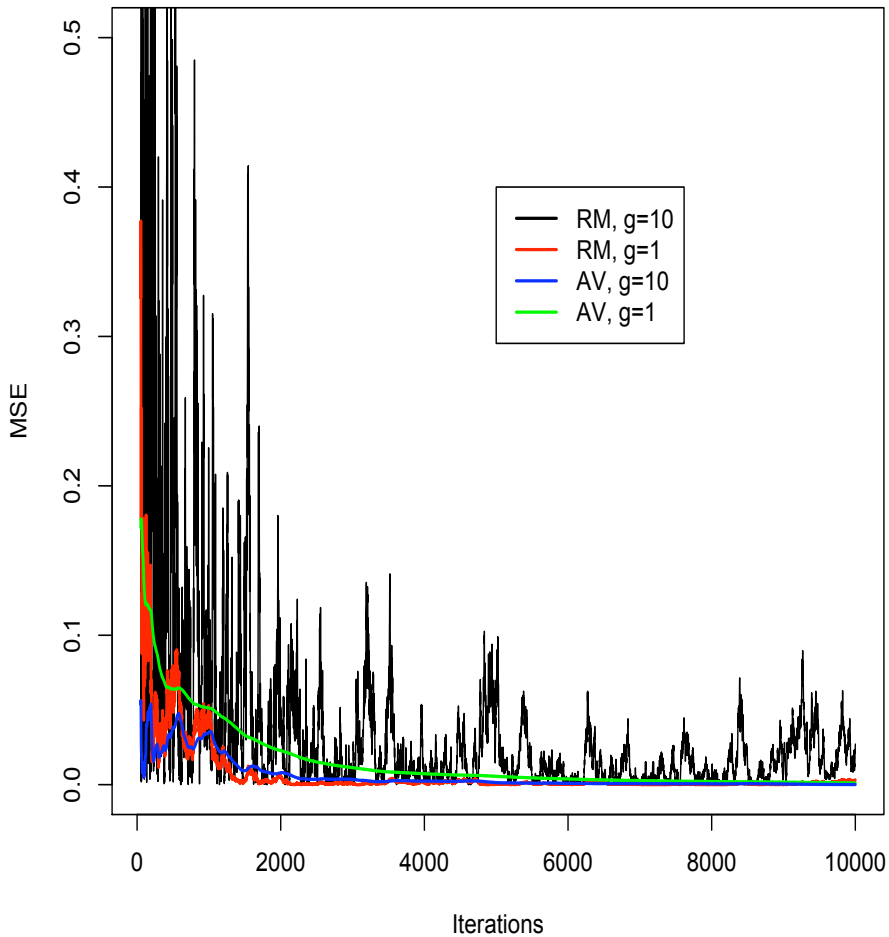
A sample with Gaussian distribution,

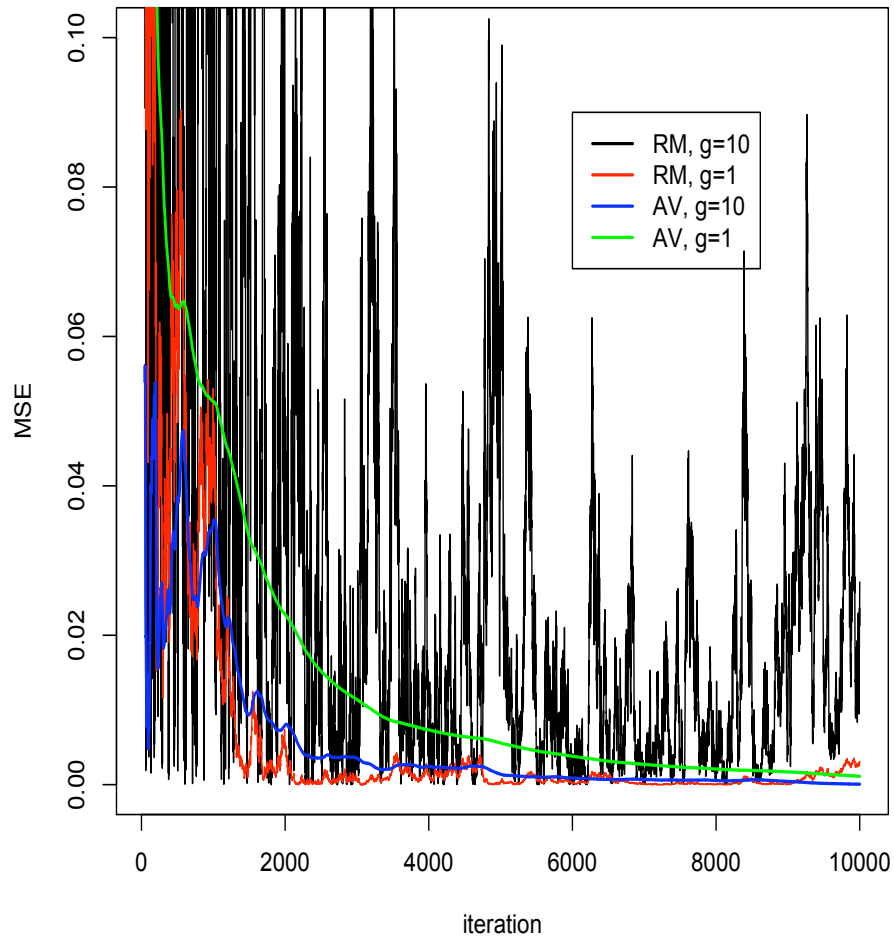
with mean $(0, 0)$ and variance $\begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix}$.



Not really, even for simple examples!!!

$$m_{n+1} = m_n + \frac{g}{n^{3/4}} \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}$$





Averaging : a magic formula

Consider now the mean of all past iterations, $\bar{m}_n = \frac{1}{n}\sum_{j=1}^{n} m_j$,

$$\begin{cases} m_{n+1} &= m_n + \gamma_n \dfrac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|} \\ \bar{m}_{n+1} &= \bar{m}_n + \dfrac{m_{n+1} - \bar{m}_n}{n+1} \end{cases}$$

Property (in $\mathbb{R}^d$)

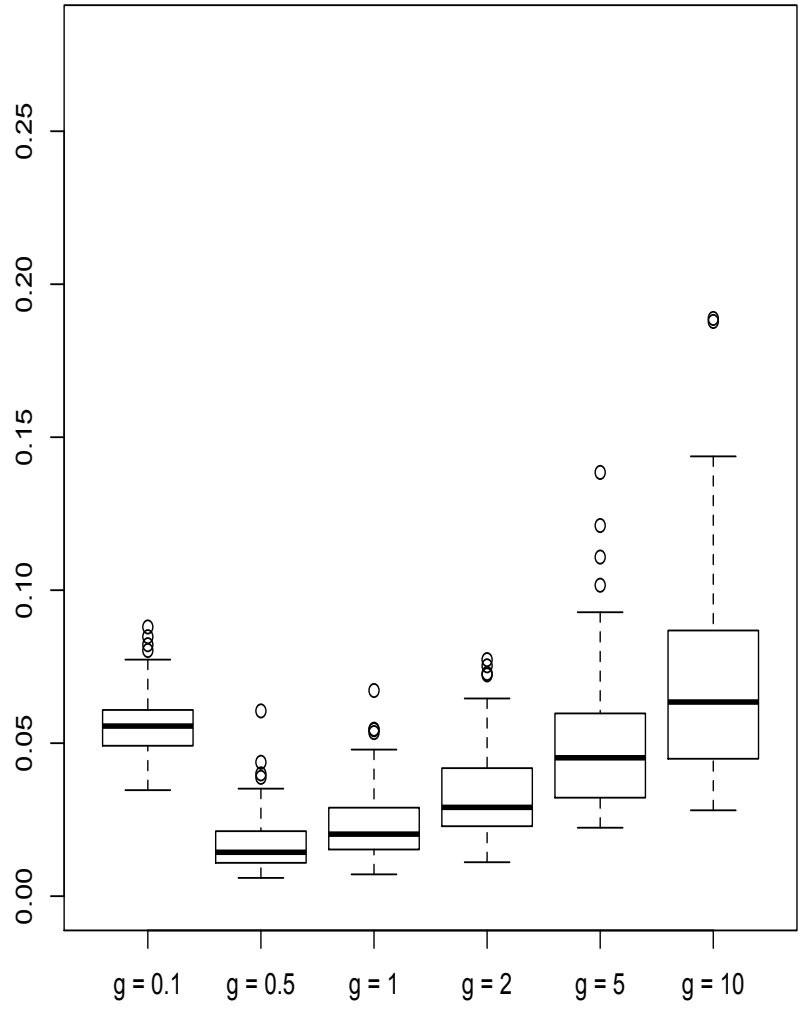
- *If* $\gamma_n = g/n^\alpha$, $0.5 < \alpha < 1$,

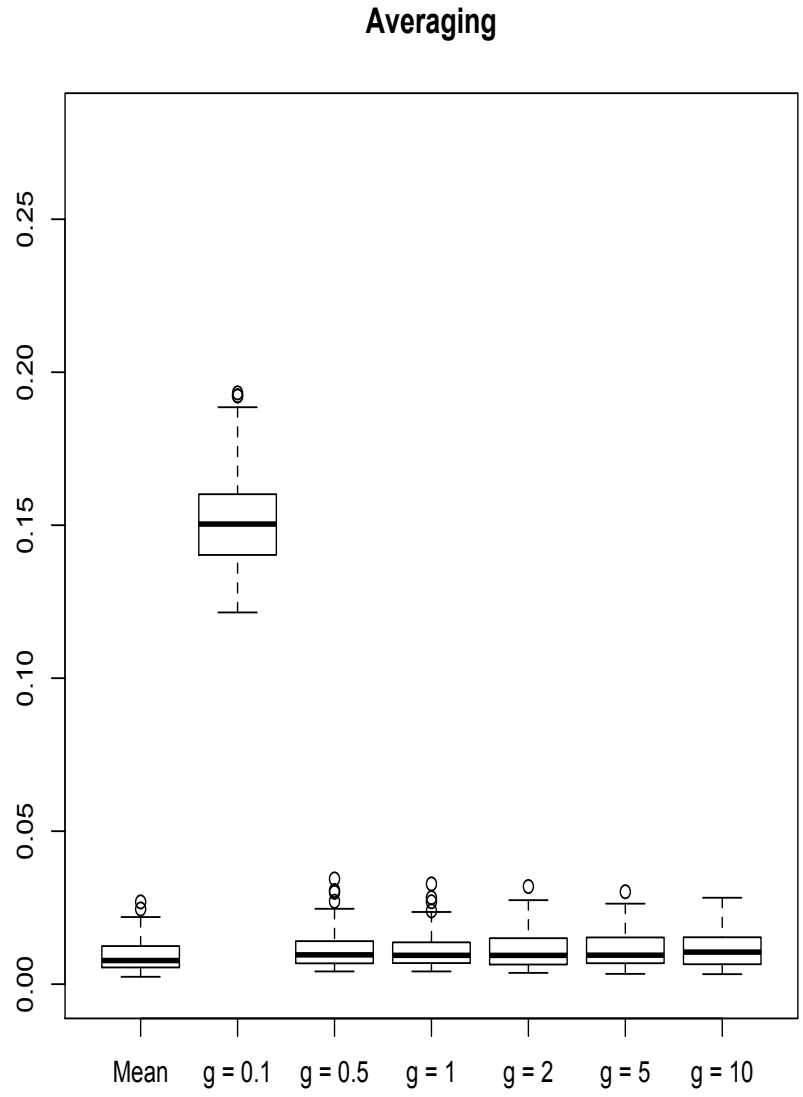
$$\sqrt{n}\,(\bar{m}_n - m) \rightsquigarrow \mathcal{N}(0, \Delta) \quad \textit{in distribution,}$$

where Δ *is the same variance matrix as for* $\widehat{m}_n$,
$\Delta = \Gamma_m^{-1} Var(S(X - m))\Gamma_m^{-1}$ *with* $S(u) = u/\|u\|$, $u \in \mathbb{R}^d$.

Proof : As in Polyak & Juditsky (1992).

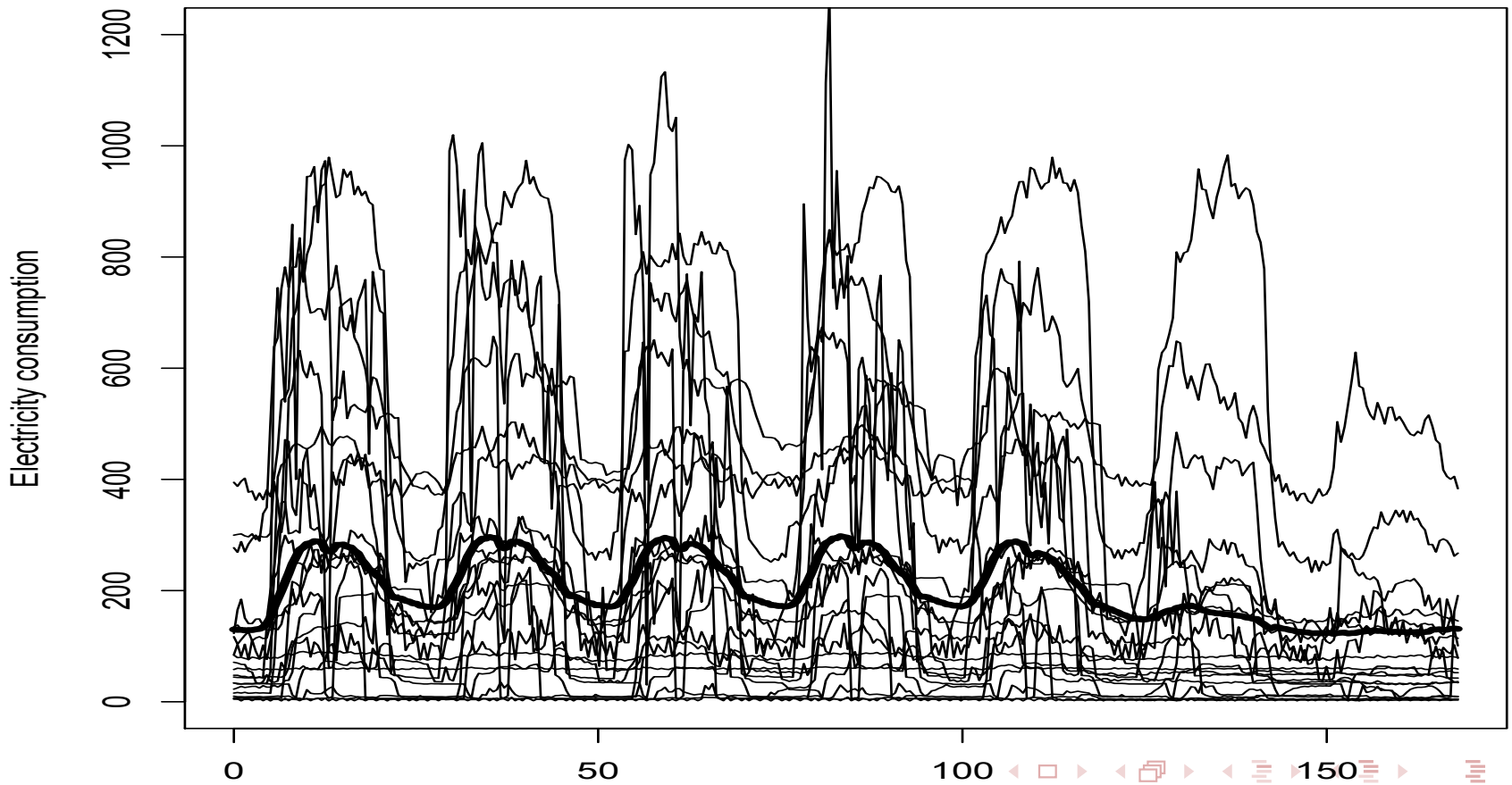
200 samples with size $n = 2000$.



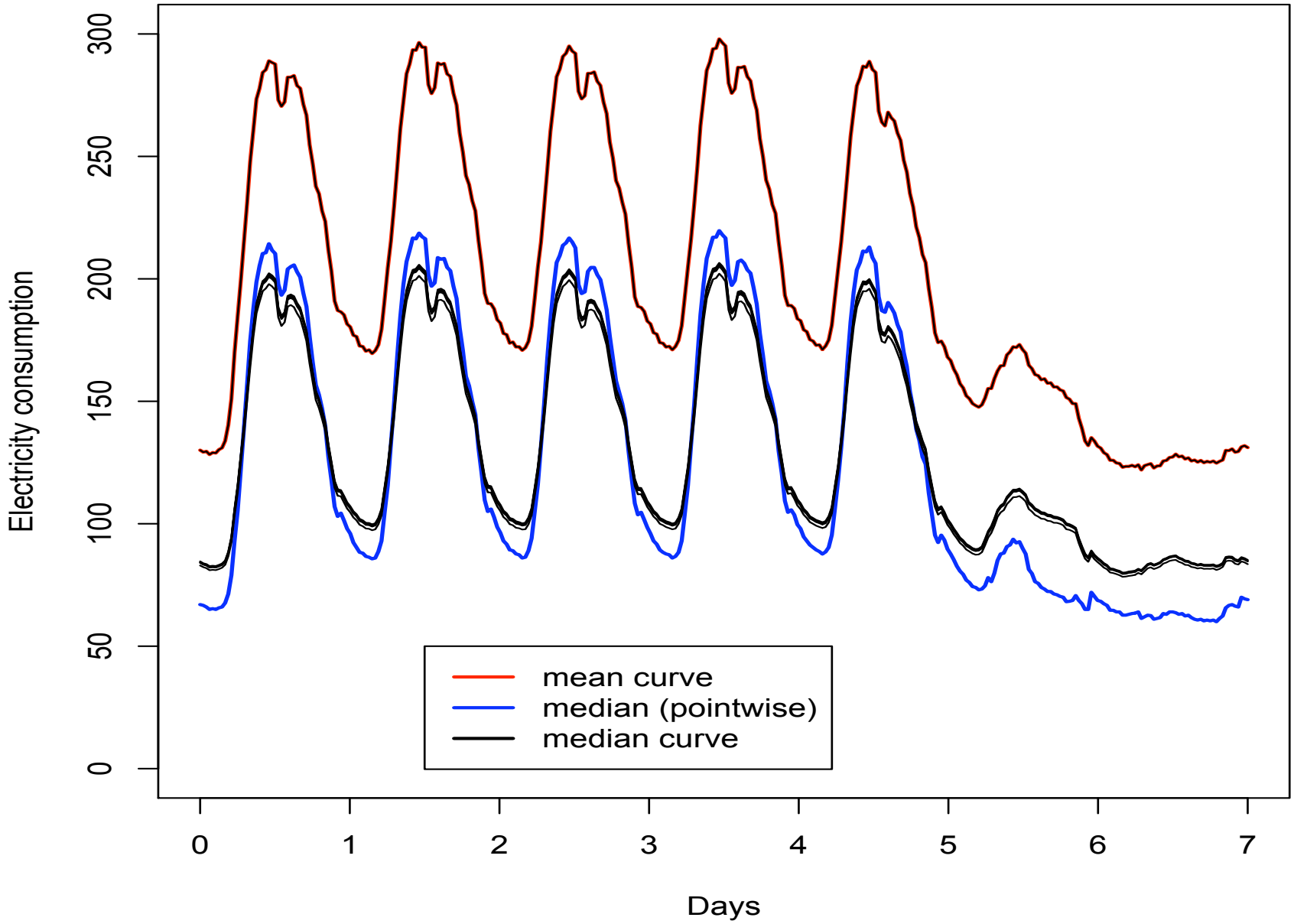


Example : electricity consumption curves

- EDF (Electricité de France) has planned to install communicating meters (30 millions). Survey sampling techniques will be used to select 300 000 meters which will provide individual electricity consumption at fine time scales.
- A first test on a population of $N = 18900$ giving electricity consumption every 30 minutes.



Example : median trajectory



Perspectives

- Averaging in Hilbert spaces H :
  still no results on nonlinear algorithms in the literature.
- Discretized noisy trajectories

$$\mathbf{Z}_n = \left(X_n(t_1^n) + \epsilon_{1n}, \ldots, X_n(t_{p_n}^n) + \epsilon_{p_n n}\right)$$

- Covariates : *conditional geometric median*

$$m(X|Z = z) = \beta_0 + \beta_1 z$$

  where Z is for example the mean consumption of the week before.
  We look for

$$\min_{(\beta_0,\beta_1)\in H\times H} \mathbb{E}\|X - (\beta_0 + \beta_1 Z)\|$$

- (robust) Clustering with medians based on $\|.\|$ (k-median).