Recursive estimation of the Median Covariation Matrix in Hilbert spaces

Antoine Godichon-Baggioni,
with Hervé Cardot and Peggy Cénac

Colloque Jeunes Probabilistes et Statisticiens 2016

18-22 Avril 2016

1. Introduction

1. Introduction

2. Estimating the geometric median
   - Definition of the median
   - The Robbins-Monro algorithm

Problematic

We want to get indicators of location and dispersion when :

Problematic

We want to get indicators of location and dispersion when :

- the sample is large,

Problematic

We want to get indicators of location and dispersion when :

- the sample is large,
- the data belong to a high dimensional space.

Problematic

We want to get indicators of location and dispersion when :

- the sample is large,
- the data belong to a high dimensional space.

For this, we want :

- a fast algorithm,

Problematic

We want to get indicators of location and dispersion when :

- the sample is large,
- the data belong to a high dimensional space.

For this, we want :

- a fast algorithm,
- which does not need to store all the data,

Problematic

We want to get indicators of location and dispersion when :

- the sample is large,
- the data belong to a high dimensional space.

For this, we want :

- a fast algorithm,
- which does not need to store all the data,
- which can be simply updated.

A simple example

The mean and the variance can be estimated recursively by

$$\overline{X}_1 = X_1,$$
$$\overline{X}_{n+1} = \overline{X}_n + \frac{1}{n+1}\left(X_{n+1} - \overline{X}_n,\right)$$
$$\overline{\Sigma}_1 = 0,$$
$$\overline{\Sigma}_{n+1} = \overline{\Sigma_n} + \frac{1}{n+1}\left(\left(X_{n+1} - \overline{X}_n\right) \otimes \left(X_{n+1} - \overline{X}_n\right) - \overline{\Sigma}_n\right).$$

Nevertheless, these indicators are not robust at all.

The geometric median in normed vector spaces

A natural generalisation of the median in a normed vector space H is given by

$$m := \arg\min_{h \in H} \mathbb{E}\left[\|X - h\| - \|X\|\right]. \tag{1}$$

The geometric median in normed vector spaces

A natural generalisation of the median in a normed vector space H is given by

$$m := \arg\min_{h \in H} \mathbb{E}\left[\|X - h\| - \|X\| \right]. \tag{1}$$

Theorem ([Kemperman, 1987])

If H is strictly convex and if the support of X is not included on a straight line, the median m is uniquely defined.

Assumptions

Let H be a separable Hilbert space (not necessarily of finite dimension)

Assumptions

(A1) The random variable X is not concentrated on a straight line : for all $h \in H$, there is $h' \in H$ such that $\langle h, h' \rangle = 0$ and

$$\mathrm{Var}\langle X, h' \rangle > 0.$$

Assumptions

Let H be a separable Hilbert space (not necessarily of finite dimension)

Assumptions

(A1) The random variable X is not concentrated on a straight line : for all $h \in H$, there is $h' \in H$ such that $\langle h, h' \rangle = 0$ and

$$\mathrm{Var}\langle X, h' \rangle > 0.$$

(A2) The random variable X is not concentrated around single points : there is a positive constant C such that for all $h \in H$:

$$\mathbb{E}\left[\frac{1}{\|X-h\|}\right] \leq C, \qquad \mathbb{E}\left[\frac{1}{\|X-h\|^2}\right] \leq C.$$

Convexity properties

Let $G : H \longrightarrow \mathbb{R}$ be the function we would like to minimize defined for all $h \in H$ by

$$G(h) := \mathbb{E}\left[\|X - h\| - \|X\|\right]. \tag{2}$$

G is a convex function.

Convexity properties

Let $G : H \longrightarrow \mathbb{R}$ be the function we would like to minimize defined for all $h \in H$ by

$$G(h) := \mathbb{E}\left[\|X - h\| - \|X\| \right]. \qquad (2)$$

G is a convex function.

Proposition

Assuming (A1) and (A2), G is twice Fréchet-differentiable and m is the unique solution of the equation

$$\Phi(h) := \nabla G(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|} \right] = 0.$$

The Robbins-Monro algorithm

Let $X_1, ..., X_n, ...$ be independent variables with the same law as X. The first estimator of the median [Cardot et al., 2013] is given recursively by

$$m_{n+1} = m_n + \gamma_n \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}, \tag{3}$$

The Robbins-Monro algorithm

Let $X_1, ..., X_n, ...$ be independent variables with the same law as X. The first estimator of the median [Cardot et al., 2013] is given recursively by

$$m_{n+1} = m_n + \gamma_n \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}, \qquad (3)$$

with m_1 chosen bounded and (γ_n) is a decreasing sequence of positive real numbers which verifies the following usual conditions

$$\sum_{n\geq 1} \gamma_n = \infty, \qquad \sum_{n\geq 1} \gamma_n^2 < \infty.$$

Rates of convergence

We suppose from now that the step sequence is of the form $\gamma_n = c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in (\frac{1}{2}, 1)$.

Theorem ([Godichon-Baggioni, 2015])

Under assumptions, for all positive integer p, there is a positive constant C_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\|m_n - m\|^{2p}\right] \leq \frac{C_p}{n^{p\alpha}}.$$

Rates of convergence

We suppose from now that the step sequence is of the form $\gamma_n = c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in (\frac{1}{2}, 1)$.

Theorem ([Godichon-Baggioni, 2015])

Under assumptions, for all positive integer p, there is a positive constant C_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\|m_n - m\|^{2p}\right] \leq \frac{C_p}{n^{p\alpha}}.$$

Note that as a particular case, we have the following rate of convergence in quadratic mean

$$\mathbb{E}\left[\|m_n - m\|^2\right] = O\left(\frac{1}{n^\alpha}\right),$$

which is the optimal rate for Robbins-Monro algorithms.

The averaged algorithm

In order to improve the convergence, we now define recursively the averaged algorithm :

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \tag{4}$$

with $\overline{m}_1 = m_1$.

The averaged algorithm

In order to improve the convergence, we now define recursively the averaged algorithm :

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \tag{4}$$

with $\overline{m}_1 = m_1$. This also can be written as

$$\overline{m}_n = \frac{1}{n}\sum_{k=1}^{n} m_k.$$

Rate of convergence

The following theorem gives the L^p rates of convergence of the averaged algorithm.

Theorem ([Godichon-Baggioni, 2015])

Under assumptions, for all integer $p \geq 1$, there is a positive constant K_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{m}_n - m\right\|^{2p}\right] \leq \frac{K_p}{n^p}.$$

Definition of the Median Covariation Matrix

We now consider the space of linear operators on H denoted by $\mathcal{S}(H)$ and equipped with the Froebenius (or Hilbert-Schmidt) inner product. More precisely, let $(e_i)_{i\in I}$ be an orthonormal basis of H and let $A, B \in \mathcal{S}(H)$,

$$\langle A, B \rangle_F = \sum_{i\in I} \langle A(e_i), B(e_i) \rangle .$$

Definition of the Median Covariation Matrix

We now consider the space of linear operators on H denoted by $\mathcal{S}(H)$ and equipped with the Froebenius (or Hilbert-Schmidt) inner product. More precisely, let $(e_i)_{i\in I}$ be an orthonormal basis of H and let $A, B \in \mathcal{S}(H)$,

$$\langle A, B\rangle_F = \sum_{i\in I} \langle A(e_i), B(e_i)\rangle .$$

Let X be a random variable taking values in H, its Median Covariation Matrix V is defined by

$$V = \arg\min_{\Gamma\in\mathcal{S}(H)} \mathbb{E}\left[\|(X-m)\otimes(X-m)-\Gamma\|_F - \|(X-m)\otimes(X-m)\|_F\right]. \tag{5}$$

Assumptions

We suppose from now that the following assumptions are fulfilled.

Assumptions

(A3) For all $\Gamma \in \mathcal{S}(H)$*, there is* $\Gamma' \in \mathcal{S}(H)$ *such that* $\langle \Gamma, \Gamma' \rangle = 0$ *and*

$$\mathrm{Var} \left\langle (X - m) \otimes (X - m), \Gamma' \right\rangle_F > 0.$$

Assumptions

We suppose from now that the following assumptions are fulfilled.

Assumptions

(A3) For all $\Gamma \in \mathcal{S}(H)$, *there is* $\Gamma' \in \mathcal{S}(H)$ *such that* $\langle \Gamma, \Gamma' \rangle = 0$ *and*

$$\mathrm{Var} \left\langle (X - m) \otimes (X - m), \Gamma' \right\rangle_F > 0.$$

(A4) There is a positive constant C *such that for all* $h \in H$ *and for all* $\Gamma \in \mathcal{S}(H)$,

$$\mathbb{E}\left[\frac{1}{\| (X-h) \otimes (X-h) - \Gamma \|_F^2} \right] \leq C.$$

The algorithms

We can both estimate the geometric median and the MCM with the help of recursive algorithms as follows :

$$
\begin{aligned}
m_{n+1} &= m_n + \gamma_n \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|}, \\
\overline{m}_{n+1} &= \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \\
V_{n+1} &= V_n + \gamma_n \frac{(X_{n+1} - \overline{m}_n) \otimes (X_{n+1} - \overline{m}_n) - V_n}{\|(X_{n+1} - \overline{m}_n) \otimes (X_{n+1} - \overline{m}_n) - V_n\|_F}, \\
\overline{V}_{n+1} &= \overline{V}_n + \frac{1}{n+1}\left(V_{n+1} - \overline{V}_n\right),
\end{aligned}
$$

with m_1, V_1 chosen bounded and $\overline{m}_1 = m_1$, $\overline{V}_1 = V_1$.

Rate of convergence of the stochastic gradient algorithm

We suppose from now that the step sequence is of the form $\gamma_n := c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$. The following theorem gives the rate of convergence in quadratic mean of the stochastic gradient algorithm.

Theorem ([Cardot and Godichon-Baggioni, 2015])

Suppose assumptions hold, let $\beta \in (\alpha, 2\alpha)$, there are positive constants C, C_β such that for all $n \geq 1$,

$$\mathbb{E}\left[\|V_n - V\|_F^2\right] \leq \frac{C}{n^\alpha},$$
$$\mathbb{E}\left[\|V_n - V\|_F^4\right] \leq \frac{C_\beta}{n^\beta}$$

Rate of convergence of the averaged algorithm

The following theorem gives the rate of convergence in quadratic mean of the averaged algorithm.

Theorem ([Cardot and Godichon-Baggioni, 2015])

Suppose assumptions hold, there is a positive constant K such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{V}_n - V\right\|_F^2\right] \leq \frac{K}{n}.$$

Application to Robust PCA

Theorem ([Kraus and Panaretos, 2012])

If the distribution of X is symmetric around m and if $\mathbb{E}\left[\|X\|^2\right] < +\infty$, the Median Covariation Matrix V and the covariance matrix $\Sigma := \mathbb{E}\left[(X - m) \otimes (X - m)\right]$ have the same eigenvectors.

Application to Robust PCA

Theorem ([Kraus and Panaretos, 2012])

If the distribution of X is symmetric around m and if $\mathbb{E}\left[\|X\|^2\right] < +\infty$, the Median Covariation Matrix V and the covariance matrix $\Sigma := \mathbb{E}\left[(X-m)\otimes(X-m)\right]$ have the same eigenvectors.

Moreover, we can estimate online the q main eigenvectors of the Median Covariation Matrix with the help of the following algorithm :

$$u_{j,n+1} = u_{j,n} + \frac{1}{n+1}\left(\overline{V}_{n+1}\frac{u_{j,n}}{\|u_{j,n}\|} - u_{j,n}\right) \quad j = 1, \dots, q,$$

combined with an orthonormalization of $u_{1,n+1}, \dots, u_{q,n+1}$.

Thank you for your attention !

Cardot, H., Cénac, P., and Godichon-Baggioni, A. (2015).
Online estimation of the geometric median in Hilbert spaces : non asymptotic confidence balls.
Accepted in the Annals of Statistics.

Cardot, H., Cénac, P., Zitt, P.-A. (2013).
Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm.
Bernoulli, 19(1) :18–43.

Cardot, H. and Godichon-Baggioni, A. (2015).
Fast Estimation of the Median Covariation Matrix with Application to Online Robust Principal Components Analysis.
arXiv preprint arXiv :1504.02852.

Godichon-Baggioni, A. (2015).
Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms : l^p and almost sure rates of convergence.
Journal of Multivariate Analysis.

Kemperman, J. H. B. (1987).
The median of a finite measure on a Banach space.
In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam.

Kraus, D. and Panaretos, V. M. (2012).
Dispersion operators and resistant second-order functional data analysis.
Biometrika, 99 :813–832.