Algorithmes stochastiques pour l'estimation robuste en grande dimension

Antoine Godichon-Baggioni [1] & Hervé Cardot [2] & Peggy Cénac [3]

[1] *INSA de Toulouse, 135 Avenue de Rangueil, 31400 Toulouse, godichon@insa-toulouse.fr*

[2,3] *Institut de Mathématiques de Bourgogne, 9 Avenue Alain Savary, 21000 Dijon, herve.cardot@u-bourgogne.fr, peggy.cenac@u-bourgogne.fr*

Résumé. La médiane géométrique, aussi appelée L^1 médiane est souvent utilisée en statistique du fait de sa robustesse. De plus, il est de plus en plus usuel de traiter de gros échantillons à valeurs dans des espaces de grande dimension. Dans ce contexte, on se concentre sur des estimateurs rapide de la médiane, qui consistent en des algorithmes de gradient stochastiques moyennés. On définit aussi un nouvel indicateur de dispersion robuste (lié à la médiane) appelé Matrice de Covariance Médiane, et on donne des algorithmes pour l'estimer. Cette matrice peut être très intéressantes pour l'Analyse en Composantes Principales Robuste. En effet, sous certaines conditions, elle a les mêmes espaces propres que la matrice de covariance, mais est moins sensible aux données atypiques.

Mots-clés. Algorithmes de gradient stochastiques, Grande dimension, Statistique robuste, Médiane Géométrique, Analyse en Composantes Principales robuste.

Abstrac. The geometric median, also called L^1-median is often used in statistics because of its robustness. Moreover, it is more and more usual to deal with large sample taking values in high dimensional spaces. In this context, we focus on a fast estimator of the median which consists in an averaged stochastic gradient algorithm. We propose to give a deep study of these estimators. We also define a new robust dispersion matrix (closely related to the median) called Median Covariation Matrix and give algorithms to estimate it. This matrix can be very interesting in robust Principal Components Analysis. Indeed, under assumptions, it has the same eigenspaces as the covariation matrix, but it is less sensitive to outliers.

Mots-clés. Stochastic gradient algorithms, High dimension, Robust statistics, Geometric Median, Robust Principal Component Analysis.

1 Introduction

La médiane géométrique est une généralisation naturelle de la médiane réelle introduite par Haldane (1948). En dimension finie, on peut faire le lien avec le problème de Fermat-Webber qui consiste à minimiser la somme des distances à des points donnés et qui est un

problème d'optimisation convexe bien connu. Beaucoup de propriétés de la médiane dans les espaces de Banach sont données par Kemperman (1987), telles que son existence, son unicité et sa robustesse. Cette dernière propriété représente l'un des principaux facteurs d'intérêt de la médiane. De plus, Chakraborty and Chaudhuri (2014) proposent une étude approfondie des estimateurs pour le cas général des espaces de Banach.

Il existe quelques estimateurs de la médiane dans la littérature, et un des plus utilisé en dimension finie est celui introduit par Vardi and Zhang (2000), qui consiste a résoudre le problème de Fermat-Webber généré par l'échantillon a l'aide de l'algorithme de Weizfeld. Cependant, cet algorithme peut être difficile à compiler lorsque l'on a de grands échantillons à valeurs dans des espaces de grandes dimensions. C'est pourquoi on s'intéresse aux estimateurs de la médiane obtenus l'aide d'algorithmes de gradient stochastiques moyennés (Cardot et al., 2013). On exhibe leurs vitesses de convergence L^p ainsi que des boules de confiances.

De plus, on s'intéresse à la Matrice de Covariance Médiane (MCM) étudiée par Kraus et Panaretos (2012). C'est un indicateur de dispersion robuste en lien avec la médiane, qui peut etre utilisé pour l'ACP robuste (Cardot and Godichon-Baggioni, 2016). En effet, si la distribution des données que l'on veut étudier est symétrique, alors la MCM a les mêmes espaces propres que la matrice de covariance. On introduit donc un algorithme capable d'estimer cet indicateur quelles que soient les tailles de l'échantillon et de l'espace. Cet algorithme consiste à estimer simultanément la médiane et la MCM à l'aide d'algorithmes de gradient stochastiques et de leurs moyennés.

2 Estimation de la médiane géométrique

2.1 Définition et hypothèses

On considère une variable aléatoire X à valeurs dans un espace de Hilbert séparable H (pas nécéssairement de dimension finie). On note $\langle ., . \rangle$ le produit scalaire et $\|.\|$ la norme associée. La médiane géométrique m de X est définie par

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

On introduit maintenant deux hypothèses:

(A1) X n'est pas concentrée sur une droite: pour tout $h \in H$, il existe $h' \in H$ tel que $\langle h, h' \rangle$ et

$$Var\left(\langle X, h' \rangle\right) > 0.$$

(A2) X n'est pas concentrée autour de points isolés: il existe une constante positive C telle que pour tout $h \in H$,

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

L'hypothèse **(A1)** assure l'unicité de la médiane (Kemperman (1987)), et l'hypothèse **(A2)** permet de donner quelques propriétés sur la fonction que l'on veut minimiser.

2.2 Les algorithmes

Soit G la fonction que l'on veut minimiser. Elle est définie pour tout $h \in H$ par

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

Sous les hypothèses, G est convexe et est deux fois différentiable. Son gradient est défini pour tout $h \in H$ par

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

et la médiane m est l'unique zéro du gradient. Cela légitimise l'utilisation d'un algorithme de type Robbins-Monro. On considère maintenant des variables aléatoires indépendantes $X_1, ..., X_n, ...$ de mêmes lois que X. On rappelle maintenant l'algorithme de type Robbins-Monro introduit par Cardot et al. (2013) et défini de manière récursive par

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

avec m_1 borné. La suite de pas (γ_n) est une suite de réels positifs, décroissante, et vérifie les conditions usuelles suivantes (voir Duflo (1997) par exemple)

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

La version moyennée de l'algorithme est définie récursivement par

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \tag{3}$$

avec $\overline{m}_1 = m_1$, ce qui peut s'écrire $\overline{m}_n = \frac{1}{n}\sum_{k=1}^{n} m_k$.

2.3 Vitesses de convergence

La forte consistance de ces algorithmes est donnée par Cardot et al. (2013). On considère maintenant une suite de pas de la forme $\gamma_n := c_\gamma n^{-\alpha}$, avec $c_\gamma > 0$ et $\alpha \in (1/2, 1)$. Le théorème suivant donne alors les vitesses de convergence de l'algorithme de type Robbins-Monro.

Théorème 2.1. *On suppose que les hypothèses **(A1)** et **(A2)** sont vérifiées. Pour tout entier $p \geq 1$, il existe une constante positive K_p telle que pour tout $n \geq 1$,*

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

Grâce aux vitesses précédentes, on peut obtenir les vitesses de convergence pour l'algorithme moyenné ainsi que des boules de confiance.

Théorème 2.2. *On suppose que les hypothèses **(A1)** et **(A2)** sont vérifiées. Pour tout entier $p \geq 1$, il existe une constante positive $K_{p'}$ telle que pour tout $n \geq 1$,*

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

De plus, pour tout $\delta \in (0,1)$, il existe un rang n_δ telle que l'on ait pour tout $n \geq n_\delta$, avec probabilité au moins $1-\delta$:

$$\|\overline{m}_n - m\| \leq \frac{4}{\lambda_{\min}}\left(\frac{2}{3n} + \frac{1}{\sqrt{n}}\right)\ln\left(\frac{4}{\delta}\right),$$

où $\lambda_{\min}$ est la plus petite valeur propre de la hessienne de G en m.

3 Estimation de la Matrice de Covariance Médiane

3.1 Définition et hypothèses

On considère maintenant un espace de Hilbert séparable H et l'espace des opérateurs linéaires de H dans H, noté $\mathcal{S}(H)$. Soit $(e_j)_{j\in J}$ une base de H, on équipe $\mathcal{S}(H)$ avec le produit scalaire suivant: soient $A, B \in \mathcal{S}(H)$,

$$\langle A, B\rangle_F = \sum_{j\in J}\langle A(e_j), B(e_j)\rangle.$$

Alors, $\mathcal{S}(H)$ est aussi un espace de Hilbert séparable et la norme associée au produit scalaire précédent, notée $\|.\|_F$, est la norme de Hilbert-Schmidt (ou Froebenius). Soit X une variable aléatoire à valeurs dans H, la Matrice de Covariance Médiane Γ_m de X est définie par

$$\Gamma_m := \arg\min_{V\in\mathcal{S}(H)} \mathbb{E}\left[\left\|(X-m)(X-m)^T - V\right\|_F - \left\|(X-m)(X-m)^T\right\|_F\right], \qquad (4)$$

où m est la médiane de X. La Matrice de Covariance Médiane Γ_m peut être vue comme la médiane géométrique de la variable aléatoire $(X-m)(X-m)^T$, et est donc robuste. De la même façon que pour la médiane, on introduit maintenant deux hypothèses:

(A3) Pour tout $V \in \mathcal{S}(H)$, il existe $V' \in \mathcal{S}(H)$ tel que $\langle V, V'\rangle_H = 0$ et

$$Var\left(\left\langle (X-m)(X-m)^T, V'\right\rangle_F\right) > 0.$$

(A4) Il existe une constante positive C telle que pour tout $h \in H$ et $V \in \mathcal{S}(H)$,

$$\mathbb{E}\left[\left\|(X-h)(X-h)^T - V\right\|_F^{-2}\right] \leq C.$$

Sous les hypothèses **(A1)** et **(A3)**, la Matrice de Covariance Médiane est bien définie et est unique.

3.2 Les algorithmes

Dans le cas où la médiane est connue, les algorithmes et leurs propriétés asymptotiques sont analogues à ceux pour l'estimation de la médiane. On suppose maintenant que m n'est pas connue et pour tout $h \in H$, soit G_h la fonction définie pour tout $V \in \mathcal{S}(H)$ par

$$G_h(V) := \mathbb{E}\left[\left\|(X-h)(X-h)^T - V\right\|_F - \left\|(X-h)(X-h)^T\right\|_F\right].$$

Ces fonctions sont différentiables et leurs gradients sont définis pour tout $V \in \mathcal{S}(H)$ par

$$\nabla G_h(V) = -\mathbb{E}\left[\frac{(X-h)(X-h)^T - V}{\|(X-h)(X-h)^T - V\|_F}\right].$$

On peut maintenant introduire un algorithme de gradient stochastique et son moyenné. Soient $X_1, ..., X_n, ...$ des variables aléatoires indépendantes de mêmes lois que X. L'algorithme de type Robbins-Monro (V_n) et son moyenné $(\overline{V}_n)$ sont définis récursivement par

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

avec $m_1 = \overline{m}_1$ et $V_1 = \overline{V}_1$ bornés.

3.3 Vitesses de convergence

On considère maintenant une suite de pas de la forme $\gamma_n := c_\gamma n^{-\alpha}$, avec $c_\gamma > 0$ et $\alpha \in (1/2, 1)$. On a alors la vitesse de convergence en moyenne quadratique suivante pour l'algorithme de gradient.

Théorème 3.1. *On suppose que les hypothèses **(A1)** à **(A4)** sont vérifiées. Il existe une constante positive K telle que pour tout $n \geq 1$,*

$$\mathbb{E}\left[\|V_n - \Gamma_m\|_F^2\right] \leq \frac{K}{n^\alpha}.$$

Finallement, le théorème suivant donne la vitesse de convergence en moyenne quadratique de l'algorithme moyenné.

Théorème 3.2. *On suppose que les hypothèses **(A1)** à **(A4)** sont vérifiées. Il existe une constante positive K telle que pour tout $n \geq 1$,*

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

Bibliographie

[1] Cardot, H., Cénac, P. and Godichon-Baggioni, A. (2015) Online estimation of the geometric median in Hilbert spaces: non asymptotic confidence balls. *A paraitre dans The Annals of Statistics*.
[2] Cardot, H., Cénac, P. and Zitt, P.-A. (2013) Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19:18–43.
[3] Cardot, H. and Godichon-Baggioni, A. (2016) Fast estimation of the median covariation matrix with application to online robust principal components analysis. *A paraitre dans TEST*.
[4] Chakraborty, A. and Chaudhuri, P. (2014) The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, 42:1203–1231.
[5] Duflo, M. (1997) *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin.
[6] Godichon-Baggioni, A. (2016) Estimating geometric median in hilbert spaces with stochastic gradient algorithms L^p and almost sure rates of convergence. *Journal of Multivariate Analysis, 146, 209-222.*
[7] Haldane, J. B. S. (1948) Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417.
[8] 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.
[9] Kraus, D. and Panaretos, V.M. (2012) Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99:813–832.
[10] Vardi, Y. and Zhang, C.H. (2000) The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426.