Algorithmes stochastiques pour la statistique robuste

Antoine Godichon-Baggioni, Stéphane Robin and Laure Sansonnet,
antoine.godichon_baggioni@upmc.fr, stephane.robin@sorbonne-universite.fr
laure.sansonnet@agrosparistech.fr
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne-Université, 75005 Paris, France

Description du sujet

Quelques mots clés: Optimisation stochastique, estimation en ligne, statistique robuste, machine learning.

Contexte: L'acquisition de données massives à valeurs dans des espaces de grande dimension est malheureusement souvent accompagnée d'une contamination de ces dernières. Dans ce contexte de données contaminées, même une faible proportion α d'individus peut corrompre des indicateurs statistiques simples tels que la moyenne ou la variance. La détection automatique de ces données atypiques n'est pas simple, et l'utilisation de techniques robustes est une alternative intéressante. Il existe de nombreux indicateurs de position robustes Small (1990). Par exemple, les moyennes tronquées Rousseeuw and Leroy (2005); Fraiman and Muniz (2001) consistent à prendre la moyenne des $(1-\alpha)n$ informations les plus centrales. Cependant, cette approche nécessite d'avoir une idée de la proportion de données contaminées et suppose que ces dernières sont nécessairement éloignées de 0. De plus, ces approches nécessitent souvent des efforts de calcul importants, bien que des méthodes aient été développées pour traiter les problèmes de dimensionnalité Cuevas et al. (2007).

On choisit ici de se concentrer davantage sur la médiane géométrique (également appelée médiane L^1 ou médiane spatiale) introduite par Haldane (1948). En effet, cet indicateur de position est connu pour avoir un point de rupture de 0.5, ce qui signifie que même si près de la moitié de l'échantillon est contaminée, on peut contrôler la divergence des estimations, contrairement à la moyenne qui a un point de rupture de 0. De plus, on peut également se pencher sur la matrice de covariance médiane, qui est l'alternative robuste à la covariance habituelle. Dans un travail récent (Godichon-Baggioni and Robin, 2022), plusieurs méthodes ont été développées pour construire des estimations robustes de la variance. Cependant, aucun résultat théorique n'est établi à ce jour.

Objectifs: Après que le candidat se soit familiarisé avec la statistique robuste et les algorithmes stochastiques, les objectifs (ambitieux) du stage sont les suivants :

1. Fournir des garanties théoriques pour les estimateurs robustes de la covariance obtenues à l'aide d'une procédure de Robbins-Monro (Godichon-Baggioni and Robin, 2022).

2. En déduire des estimateurs robustes en ligne de la variance (en s'appuyant sur Cardot and Godichon-Baggioni (2015)).

3. Obtenir des garanties théoriques pour ces estimations.

4. Appliquer les méthodes proposées pour la détection en ligne des valeurs aberrantes (Rousseeuw and Driessen, 1999).

Profil recherché: Le candidat doit être en formation de M2 ou équivalent en Statistique et/ou Machine Learning.

Lieu du stage: Sorbonne Université, LPSM, 4 place Jussieu, 75005 Paris
Durée: 4 à 6 mois
Gratification: environ 550 euros par mois
Projet de thèse: Ce stage peut déboucher sur une thèse. Une fois les résultats théoriques obtenus pour les estimateurs robuste de la variance, beaucoup d'applications seraient possibles. Ces applications seraient choisies en fonction des préférences du ou de la candidat(e).

References

Cardot, H. and Godichon-Baggioni, A. (2015). Fast estimation of the median covariation matrix with application to online robust principal components analysis. *TEST*, pages 1–20.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *TEST*, 10:419–440.

Godichon-Baggioni, A. and Robin, S. (2022). A robust model-based clustering based on the geometric median and the median covariation matrix. *arXiv preprint arXiv:2211.08131*.

Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*. John wiley & sons.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277.