Université de Montpellier, France

Projet HMMA307 - M2 MIND

Utilisation de la médiane géométrique et de l'estimation robuste dans des espaces de Banach

Hanna Bacave

Novembre 2020

Table des matières

Introduction

Dans ce document, nous allons présenter la façon dont nous avons implémenté, grâce au logiciel Python, les travaux effectués par Stanislav Minsker *Geometric median and robust estimation in Banach spaces* [0]. Après une rapide présentation de ce qu'est l'estimateur médian, nous nous interesserons plus particulièrement aux travaux de Stanislav Minsker, avant de réaliser l'implémentation de ces recherches.

1 Présentation de la régression quantile

1.1 Médiane

A partir de la ressource [1], nous allons introduire la définition de la médiane.

Définition 1. *Soit $y_1, ..., y_n \in \mathbb{R}$, on définit la médiane par :*

$$Med_n(y_1, ..., y_n) \in \arg\min_{\mu \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |y_i - \mu|.$$

En pratique, on utilise plutôt une écriture de la médiane comme suit :

$$Med_n(y_1, ..., y_n) = \begin{cases} y_{\lfloor \frac{n}{2} \rfloor - 1} \ si \ n \ est \ pair \\ \frac{y_{\lfloor \frac{n}{2} \rfloor} + y_{\lfloor \frac{n}{2} \rfloor - 1}}{2} \ si \ n \ est \ impair \end{cases} .$$

1.2 Quantile

A partir de la ressource [1], nous allons introduire la définition du quantile.

Définition 2. *Le quantile de niveau α est défini par :*

$$\forall \alpha \in]0, 1], \ q_\alpha(y_1, ..., y_n) = \inf\{t \in \mathbb{R} : F_n(t) \geq \alpha\}.$$

où

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y_i \leq t\}}.$$

Remarque 1. *Le quantile de niveau $\alpha = 0.5$ est la médiane.*

Définition 3. *On définit la fonction de perte de niveau α par :*

$$l_\alpha : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \begin{cases} -(1-\alpha)x \ si \ x \geq 0 \\ \alpha x \ si \ x \leq 0 \end{cases} \end{array} .$$

1.3 Régression quantile

On dispose de
— $y_1, ..., y_n \in \mathbb{R}$ observations,
— $x_1, ..., x_n \in \mathbb{R}^p$ variables explicatives.

Définition 4. *Soit $\alpha \in]0,1[$, on appelle régression quantile les coefficients :*

$$\beta^\alpha \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} l_\alpha(y_i - x_i^T \beta),$$

où

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{nxp}.$$

Remarque 2. *La régression quantile est efficace sur des jeux de données :*
— *ayant une distribution non symétrique par rapport à la moyenne*
— *ayant une nature hétéroscédastique.*
En fait, la régression quantile est moins sensible aux points aberrants que d'autre régression.

Exemple 1. *Après avoir simulé des données dont la représentation est non symétrique par rapport à la moyenne et portant des points abérrants, nous allons comparer la régression médiane avec une régression linéaire présenté dans la Figure 1. On observe que la régression linéaire accorde plus de poids aux valeurs extrêmes ce qui lui confère un résultat moins efficace que pour la regression quantile.*

FIGURE 1 – Comparaison de la régression linéaire avec la régression quantile sur un jeu de données simples