HMMA 307 : Modèle linéaire avancés

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

Hanna Bacave
Geometric median estimation

Université de Montpellier

Table of Contents

1. Régression quantile
2. Régression quantile en grande dimension avec des données sparses

Table of Contents

1. Régression quantile

2. Régression quantile en grande dimension avec des données sparses

Définitions

Médiane

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

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

Rq. En pratique, on utilise des façons plus simples pour calculer la médiane, mais celles-ci ne sont pas adaptées aux dimensions supérieures à 1.

Definitions

- $y_1, \ldots, y_n \in \mathbb{R}$ observations,
- $x_1, \ldots, x_n \in \mathbb{R}^p$ variables explicatives.

Régression quantile

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

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

où

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

Première comparaison - Régression quantile et régression linéaire

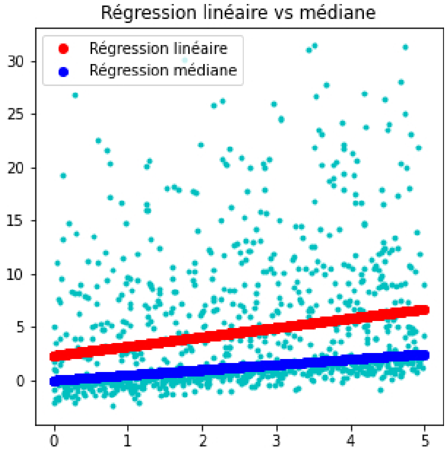

Table of Contents

1. Régression quantile
2. Régression quantile en grande dimension avec des données sparses

Présentation de la procédure - 1

Présentation du problème

On voudrait résoudre le problème d'optimisation [2] suivant :

$$Y_j = \lambda_0^T x_j + \epsilon_j$$

où

- ϵ_j est un vecteur iid de moyenne 0, mais n'est pas gaussien ;
- λ_0 est un vecteur sparse, c'est-à-dire que la dimension s du support de λ_0 est très nettement inférieure à D ;
- $D >> n$

Présentation de la procédure - 2

1. On commence par définir :
   - $t > 0$ fixé,
   - $k = 3.5t + 1$
   - $m = \frac{n}{k}$

   Pour $1 \leq l \leq k$, on définit $G_l = \{(l-1)m + 1, ..., lm\}$ et

   $$\mathbb{X}_l = (x_{j_1}|...|x_{j_m}), \text{ où, } j_i = (l-1)m + i \in G_l.$$

2. On calcule :

   $$\hat{\lambda}_\epsilon^l = \arg\min_{\lambda \in \mathbb{R}^D} \left[\frac{1}{|G_l|} \sum_{j \in G_l} (Y_j - \lambda^T x_j)^2 + \epsilon \|\lambda\|_1 \right]$$

   .

3. Enfin, on prend :

   $$\hat{\lambda}_\epsilon^* = med(\hat{\lambda}_\epsilon^1, ..., \hat{\lambda}_\epsilon^k),$$

Comparaisons des estimateurs

Méthode utilisée

On utilise le critère du R^2 pour comparer les estimateurs.

Figure: Comparaisons entre l'estimateur "Lasso médian" et d'autres estimateurs.

```
In [533]: runfile('C:/Users/hbaca/Desktop/geometric_median_estimation/code/
mediane_geom_et_comparaisons.py', wdir='C:/Users/hbaca/Desktop/
geometric_median_estimation/code')
Le R^2 de la régression quantile est 0.891967
Le R^2 de la régression quantile est 0.891967
Le R^2 de la régression lasso est de 0.999793

In [534]: runfile('C:/Users/hbaca/Desktop/geometric_median_estimation/code/
mediane_geom_et_comparaisons.py', wdir='C:/Users/hbaca/Desktop/
geometric_median_estimation/code')
Le R^2 de la régression quantile est 0.811261
Le R^2 de la régression quantile est 0.811261
Le R^2 de la régression élastic-net est de 0.999724

In [535]: runfile('C:/Users/hbaca/Desktop/geometric_median_estimation/code/
mediane_geom_et_comparaisons.py', wdir='C:/Users/hbaca/Desktop/
geometric_median_estimation/code')
Le R^2 de la régression quantile est 0.814253
Le R^2 de la régression quantile est 0.814253
Le R^2 de la régression linéaire est de 1.000000
```

Conclusion

Les principales difficultés rencontrées

- Compréhension du modèle ;
- Simulation des variables ;
- Utilisation de la médiane géométrique par le biais de la régression quantile.

Pour aller plus loin

- Comparaisons des erreurs sur des histogrammes par le biais de la validation croisée ;
- Optimisation du code.

Bibliographie

[1] Joseph Salmon, *Modéle linéaire avancé : Régression Quantile*, 2019, http://josephsalmon.eu/enseignement/Montpellier/HMMA307/RegressionQuantile.pdf ;
[2] Stanislav Minsker, *Geometric median and robust estimation in Banach spaces*, 2015, https://arxiv.org/pdf/1308.1334.pdf.