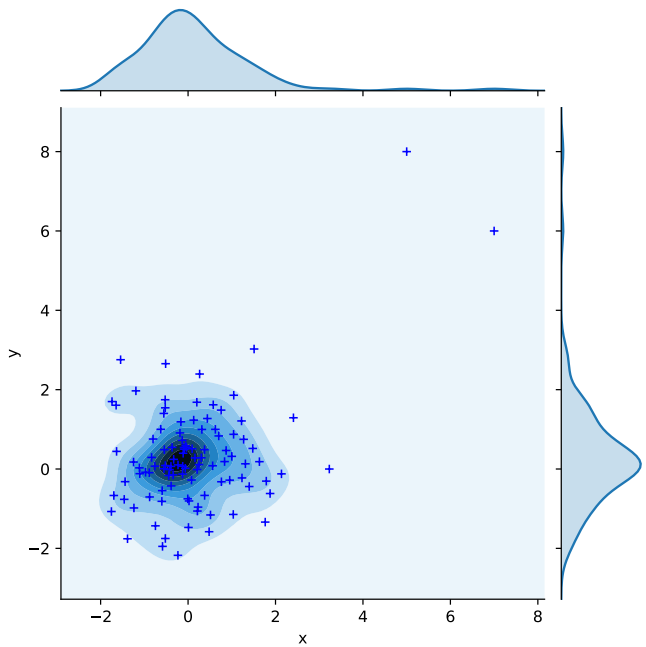
Utilisation de la Fonction d'Influence pour la Robustesse et concentration de M-estimateurs

Timothée Mathieu
Directeurs: Matthieu Lerasle, Guillaume Lecué (ENSAE)

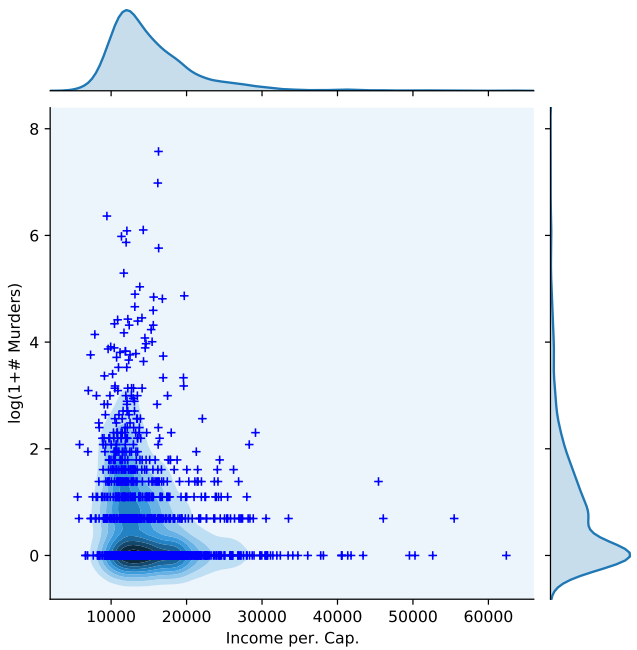
Outline

Robustesse : étude des effets que des petits changements sur les hypothèses peuvent avoir sur le résultat et des moyens de se prémunir des effets indésirables.

Exemples de "changements sur les hypothèses" :



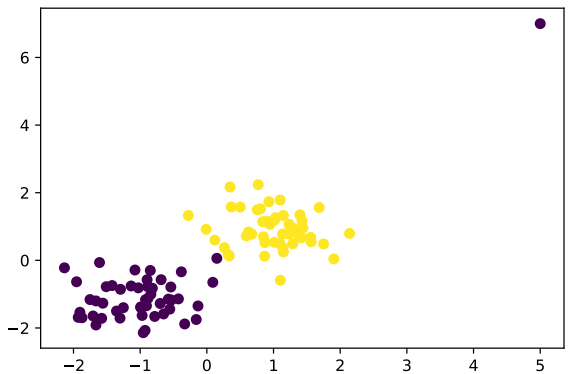
(a) 2D dataset



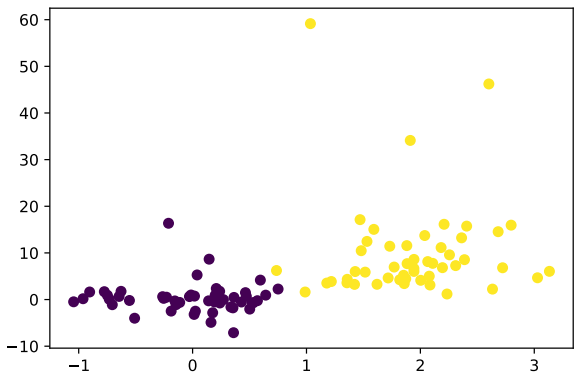
(b) Income VS # murders US

Robustesse : étude des effets que des petits changements sur les hypothèses peuvent avoir sur le résultat et des moyens de se prémunir des effets indésirables.

Exemples de "changements sur les hypothèses" :



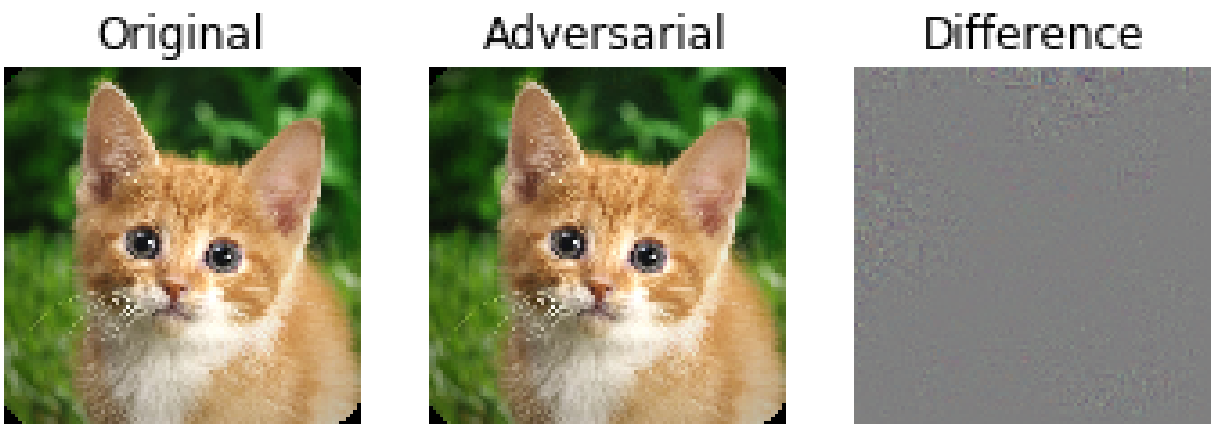
(h) Outliers en classification



(i) Heavy tail

Robustesse : étude des effets que des petits changements sur les hypothèses peuvent avoir sur le résultat et des moyens de se prémunir des effets indésirables.

Exemples de “changements sur les hypothèses” :



(o) Outliers en Réseaux de Neurones. Un chat ou un tableau d’affichage ?

Problème et estimateurs

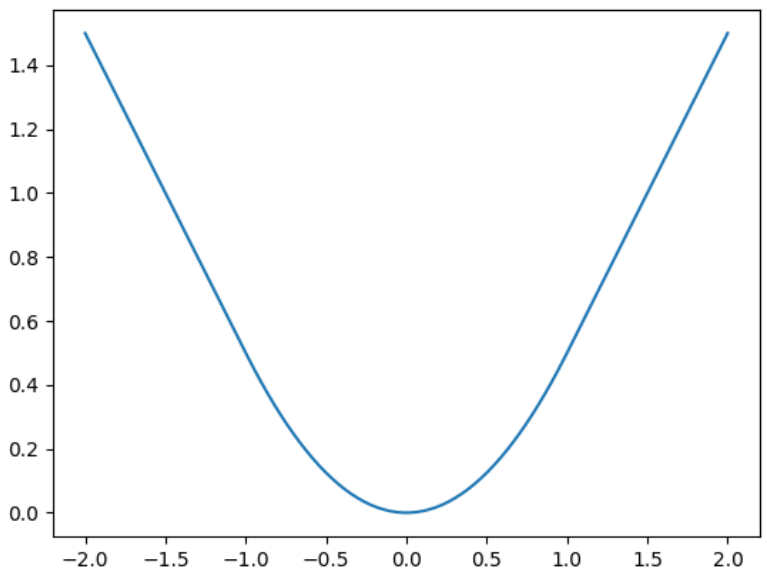
On veut estimer la moyenne $\mathbb{E}[X]$ d'une variable aléatoire X.

$$\overline{X} \in \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2$$

$\overline{X}$ n'est pas robuste. On va enlever du poids aux valeurs extrèmes [Hub64].

$$\mu \in \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho(X_i - \theta).$$

ρ :



$$\mu \in \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho(X_i - \theta).$$

μ est appellé M-estimateur. Formulation alternative avec $\psi = \rho'$,

$$\frac{1}{n} \sum_{i=1}^{n} \psi(X_i - \mu) = 0$$

Exemple:

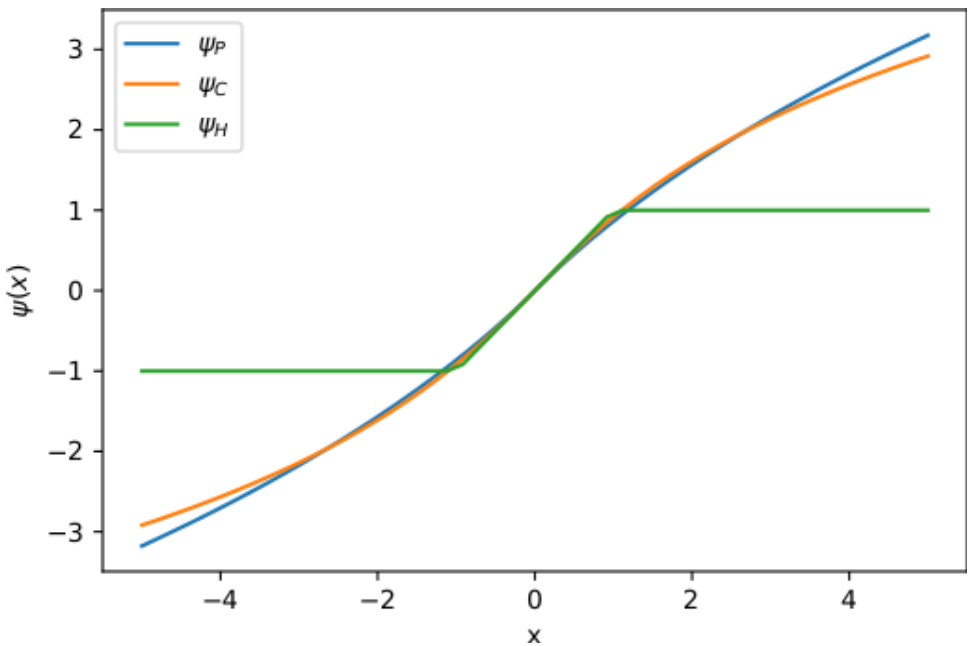
- $\psi(x) = x$:

$$\mu = \overline{X}_n$$

- $\psi(x) = \text{sign}(x)$:

$$\mu = \textit{median}(X_i)$$

- $\psi_H(x) = (-\beta \vee x) \wedge \beta$ ...



Estimateurs et fonctionelle.

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i-\mu)=0.$$

Pour étudier μ, on définit T une fonctionnelle sur les densités de proba

$$\mathbb{E}_P[\psi(X-T(P))]=0$$

On remarque que $\mu = T(\widehat{P}_n)$ et

$$\mu = T(\widehat{P}_n) \xrightarrow[n\to\infty]{proba} T(P)$$

On va ensuite étudier la fonction T grâce à sa dérivée.

L'outil utilisé

Fonction d'influence [HR09]

$$IF(x, P, T) = \lim_{\varepsilon \to 0} \frac{T(P) - T((1-\varepsilon)P + \varepsilon \delta_x)}{\varepsilon}.$$

fonction de $\mathbb{R}$ dans $\mathbb{R}$.
Prend une forme simple dans le cas M-estimateur.

$$IF(x, P, T) \propto -\psi(x - T(P))$$

Idée derrière IF : un Taylor fonctionel

$$T(Q) = T(P) + \int IF(x, P, T)[Q(dx) - P(dx)] + ...$$

Si on maitrise IF, on maitrise $|T(P) - T(Q)|$.

$$T(\widehat{P}_N) = T(P) + \frac{1}{n} \sum_{i=1}^{n} IF(X_i, P, T) + ...$$

Formalisation de la robustesse.

Nous voulons les mêmes garanties que dans le cas Gaussien quand les données sont heavy tailed. Typiquement en terme de concentration [DLLO16, Cat12].
Idée : Pour X gaussien, $\forall t > 0$

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| > \sigma\sqrt{\frac{t}{n}} \right) \leq e^{-t/2} \tag{1}$$

Pour $Var(X) < \infty$, $\forall t > 0$

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| > \sigma\sqrt{\frac{t}{n}} \right) \leq \frac{1}{t} \tag{2}$$

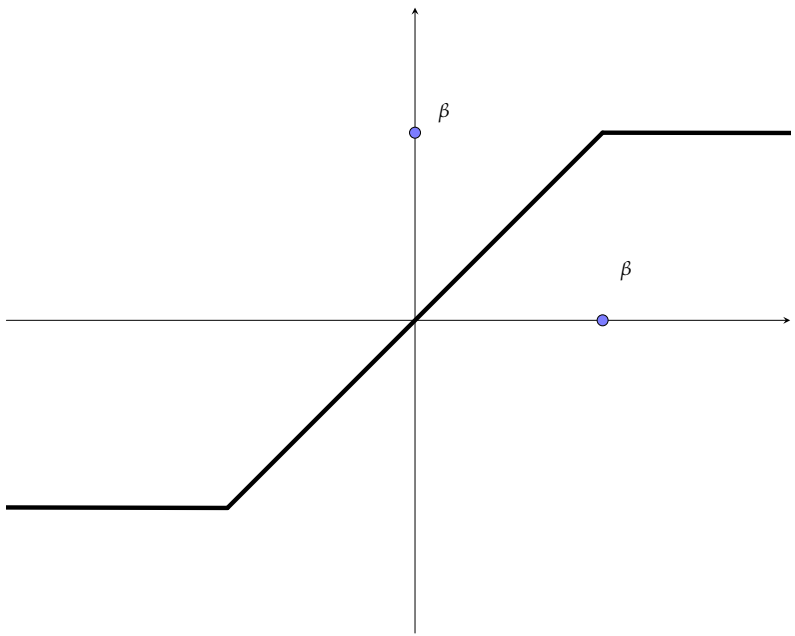
Ces inégalités sont optimal (modulo les constantes).
On veut une vitesse similaire à (1) dans le context de (2).

Concentration de M-estimateurs

Théorème informel

$$\mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big) \simeq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i - T(P))\right| > \lambda\right)$$

Soit $\beta > 0$, et ψ :



$$\text{Hoeffding :} \quad \mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big) \simeq e^{-2n\lambda^2/\beta^2}$$

Théorème

Hypothèses: ψ croissante, impaire et concave sur $\mathbb{R}_+$. Suppose que $\exists \gamma, \beta > 0$ tel que

$$\gamma \mathbb{1}\{|x| \leq \beta\} \leq \psi'(x) \leq 1$$

Alors,

- *Pour tout $\lambda > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i - T(P))\right| > 3\lambda\right) \leq \mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big).$$

- *Si de plus, $V = \mathbb{E}[\psi(|X - T(P)|)^2] \leq \psi(\beta/2)^2/2 < \infty$, alors, pour tout $\lambda \in (0, \beta/2)$,*

$$\mathbb{P}\Big(\Big|T(\widehat{P}_n) - T(P)\Big| > \lambda\Big) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\psi(X_i - T(P))\right| > \frac{\lambda\gamma}{4}\right) + e^{-n\gamma^2/8}.$$

Cas de l'estimateur de Huber : Alors, par inégalité Bernstein, si $t \leq n/8$,

$$\mathbb{P}\left(\left| T(\widehat{P}_n) - T(P) \right| > 4\sqrt{\frac{2V_H t}{n}} + 4\frac{\beta t}{n} \right) \leq 2e^{-t} + e^{-n/8}. \qquad (3)$$

où

$$V_H = Var(\psi(X - T(P))) = \mathbb{E}[\beta^2 \wedge (X - T(P))^2]$$

est finie même si $Var(X) = \infty$!
Si $\beta = \sqrt{V_H}$ on obtient une concentration sous-gaussienne autours de $T(P)$.

Deuxième caractérisation de la robustesse

Si d est une distance entre probabilités (TV, Wasserstein, Prokhorov...).
Robustesse $=$ une continuité de T:

$$d(P,Q) \leq \varepsilon \quad \Rightarrow \quad |T(P) - T(Q)| \leq \varepsilon$$

Example de corruption:

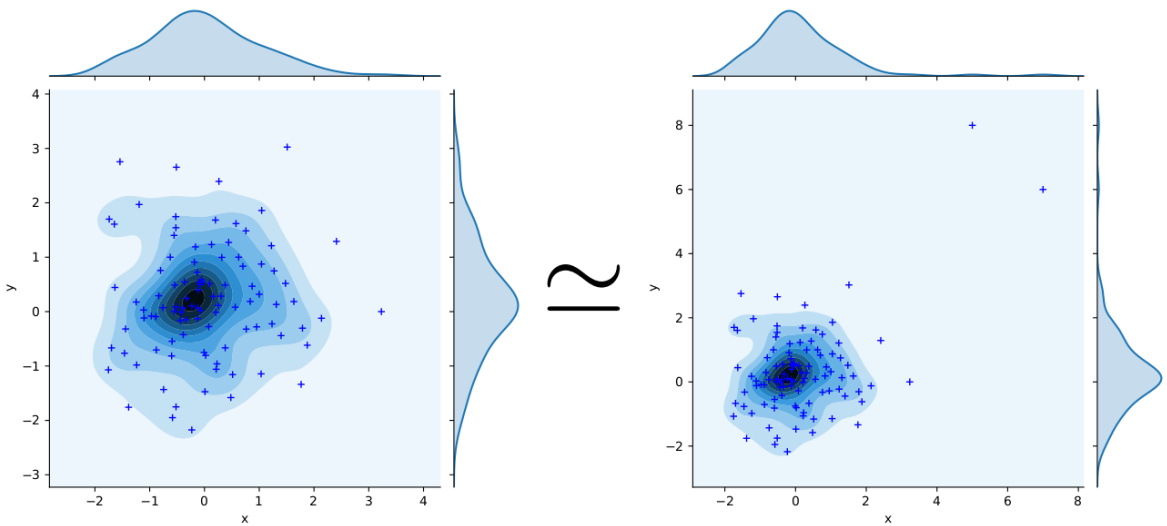


Figure 1: Corruption TV

Pour tout $\varepsilon \in (0,1)$,

$$TV((1-\varepsilon)X + \varepsilon Y, X) \leq \varepsilon$$

Deuxième caractérisation de la robustesse

Si d est une distance entre probabilités (TV, Wasserstein, Prokhorov...).
Robustesse $=$ une continuité de T:

$$d(P, Q) \leq \varepsilon \quad \Rightarrow \quad |T(P) - T(Q)| \leq \varepsilon$$

Example de corruption:

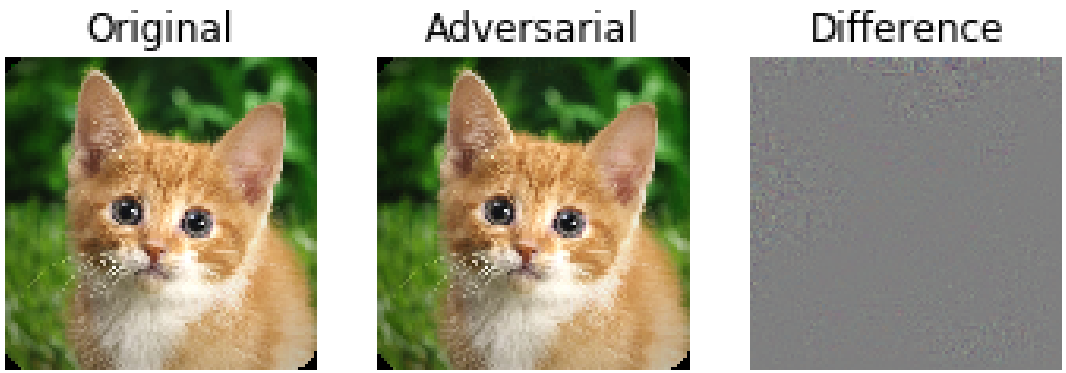


Figure 2: Corruption Wasserstein

$TV(\delta_x, \delta_y) = \mathbb{1}\{x \neq y\}$, pas de notion de distance.
Alors que $Wass(\delta_x, \delta_y) \leq \|x - y\|$.

Continuité de T

On définit la distance suivante entre probabilités (tirée du transport optimal):

$$W_\psi(P,Q) = \sup_{h \preceq \psi} \left\{ \int h(x) dP(x) - \int h(x) dQ(x) \right\}. \tag{4}$$

où $h \preceq \psi$ si et seulement si

$$\forall x, y, \quad h(x) - h(y) \leq \psi(|x-y|)$$

Exemple

- Si $\psi(x) = x$, $W_\psi = W_1$ Wasserstein.
- Si $\psi(x) = sign(x)$, $W_\psi = TV$.

Théorème

Pour toute distribution d'outliers H_t, pour toute distribution P,

$$W_\psi((1-t)P + tH_t, P) \xrightarrow[t\to 0]{} 0 \quad \textit{ssi} \quad t\mathbb{E}_{H_t}[\psi(|O_t|)] \xrightarrow[t\to 0]{} 0$$

Condition sur l'amplitude des outliers pour que $(1-t)P + tH$ converge vers P.

Théorème

T est continue par rapport à W_ψ. i.e.

$$|T(P) - T(Q)| \xrightarrow[W_\psi(P,Q)\to 0]{} 0$$

Application de cette continuité :

Théorème

Soit $X_1, \dots, X_n$ contenant k_n et $n - k_n$ i.i.d variables aléatoires $\sim P$, si les outliers vérifient

$$\mathbb{E}[\psi(|O_n|)] \leq d_n,$$

Alors

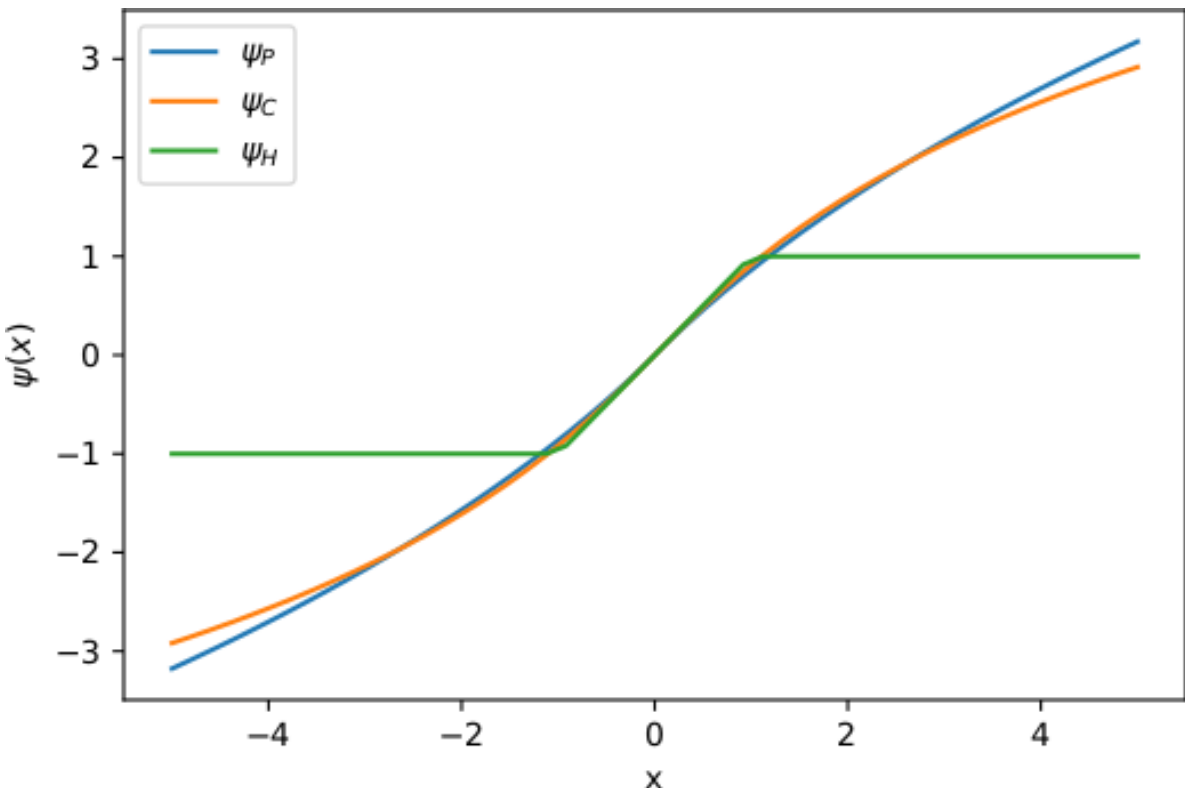
$$\frac{k_n d_n}{n} \xrightarrow{n\to\infty} 0 \quad \Rightarrow \quad \left| T\left(\widehat{P}_n\right) - T(P) \right| \xrightarrow{n\to\infty} 0,$$

Condition sur amplitude des outliers pour converger.

Illustration

On étudie trois M-estimateurs.

- Huber estimateur : ψ borné
- Catoni estimateur : ψ logarithmique à l'infini
- Polynomial estimateur : ψ en $x^{1/p}$ à l'infini, $p > 1$.



Selon le théorème précédent:

- Huber estimateur : converge quelques soient les outliers
- Catoni estimateur : converge si les outliers sont $o(e^n)$
- Polynomial estimateur : converge si les outliers sont $o(n^p)$ (ici $p = 3$).

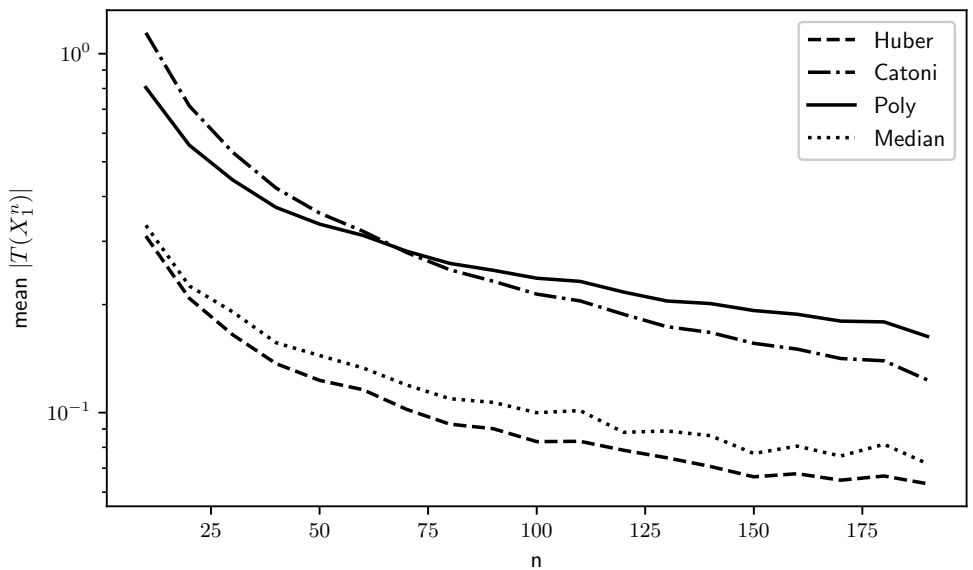


Figure 3: X Gaussien et outliers situés en n^2

Selon le théorème précédent:

- Huber estimateur : converge quelques soient les outliers
- Catoni estimateur : converge si les outliers sont $o(e^n)$
- Polynomial estimateur : converge si les outliers sont $o(n^p)$ (ici $p = 3$).

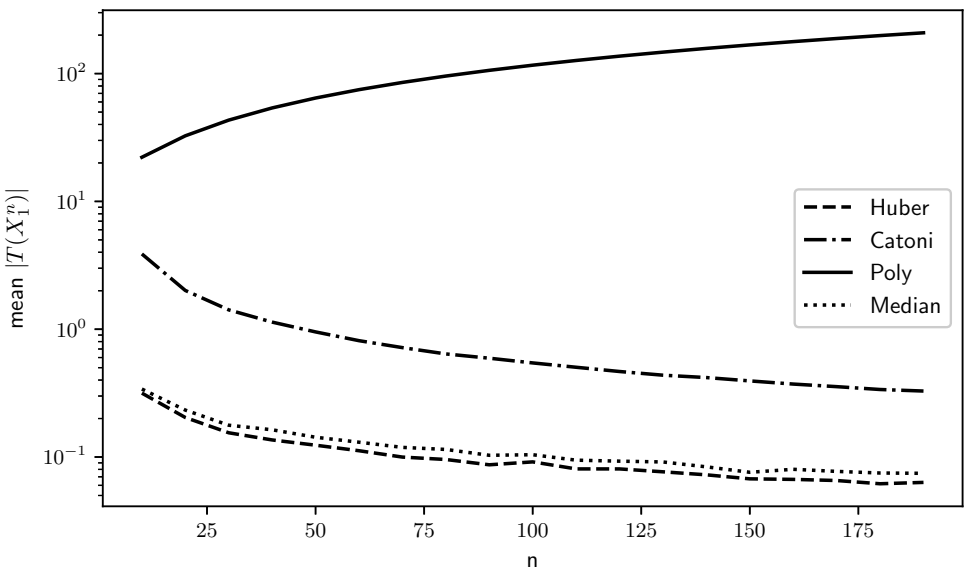


Figure 4: X Gaussien et outliers situés en n^5

Selon le théorème précédent:

- Huber estimateur : converge quelques soient les outliers
- Catoni estimateur : converge si les outliers sont $o(e^n)$
- Polynomial estimateur : converge si les outliers sont $o(n^p)$ (ici $p = 3$).

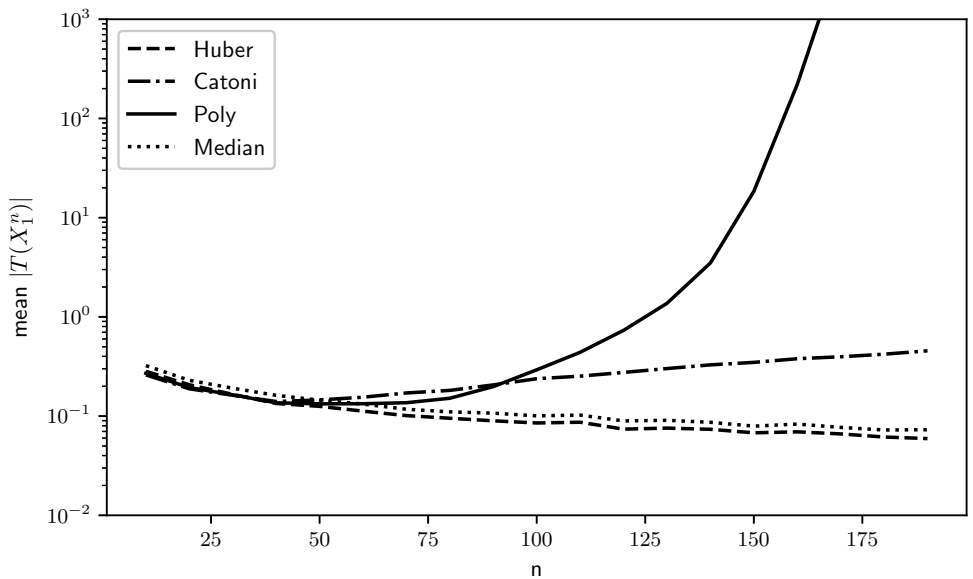


Figure 5: X Gaussien et outliers situés en $exp(n)$

Conclusion et perspectives

La fonction d'influence permet de déduire beaucoup sur l'estimateur.
Peut être étendu au multidim (fait dans l'article).

Perspectives

- Maitrise du biais $|T(P) - \mathbb{E}[X]|$
- choix du paramètre β qui est souvent le paramètre qui mesure à quel point l'estimateur est robuste.
- Application à la médiane des moyenne : quand est-il intéressant de faire des blocs ?


Olivier Catoni.
Challenging the empirical mean and empirical variance: A deviation study.
Ann. Inst. H. Poincaré Probab. Statist., 48(4):1148–1185, 11 2012.


Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira.
Sub-gaussian mean estimators.
The Annals of Statistics, 44(6):2695–2725, 2016.


Peter J Huber and Elvezio M Ronchetti.
Robust statistics; 2nd ed.
Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2009.


Peter J. Huber.
Robust estimation of a location parameter.
Ann. Math. Statist., 35(1):73–101, 03 1964.


Timothée Mathieu.
Robustness to outliers and concentration of M-estimators by means of influence function.
To appear on arxiv, 2020 ?