Groupe de Travail sur la robustesse en statistique

Yannick Baraud

7 novembre 2017

1 Le problème

Le problème de la robustesse, qui est essentiel en statistique, a été très en vogue dans les années 60 et 70 puis est tombé un peu en désuétude avant de connaitre un regain de popularité ces dernières années sous l'émergence des données en grande dimension. De manière informelle, on pourrait mettre sous le terme de *robustesse* tout ce qui concerne la stabilité des estimateurs lorsque l'on s'écarte un peu du modèle statistique que l'on a choisi.

L'exemple le plus connu est celui de l'estimation de la moyenne d'un n-échantillon gaussien $X_1, \ldots, X_n$ par sa moyenne empirique $\overline{X}_n$. Il suffit alors d'une seule valeur aberrante (très grande) au sein de l'échantillon pour que $\overline{X}_n$ fournisse une très mauvaise approximation de la vraie moyenne. En revanche, le problème disparait lorsque l'on remplace la moyenne empirique par la médiane empirique. La présence d'une valeur aberrante ne modifiera que très peu l'estimateur. La médiane empirique est robuste alors que la moyenne empirique ne l'est pas.

Un autre exemple d'estimateur très peu robuste est celui de l'estimation du paramètre θ de la loi uniforme sur $[0, \theta]$ par le maximum des observations. Même si je dispose d'un milliard de données i.i.d. qui suivent réellement une loi uniforme, disons $U([0, 1])$, il suffit d'une donnée qui vaut 100 pour que mon estimation du paramètre θ vale 100, ce qui est catastrophique. Bien entendu, il me suffirait de représenter mes données pour voir qu'il existe dans mon échantillon une donnée aberrante mais les choses se compliquent en grande dimension où il est bien plus difficile de visualiser les données en général. Dans le cas de la loi uniforme, nous pourrions de nouveau nous baser sur la médiane (en fait, 2 fois la médiane) pour estimer le paramètre θ de manière plus robuste mais l'estimateur ainsi construit ne converge qu'à vitesse $1/\sqrt{n}$ et non plus à la vitesse optimale $1/n$ comme l'estimateur précédent et l'on perd donc en qualité d'estimation. Cela n'est pas très satisfaisant

car l'on cherche à avoir des procédures qui soient à la fois robustes mais aussi optimales (ou quasi optimales) lorsque tout se passe bien.

A titre d'illustration prenons le cas de l'estimation du paramètre de translation θ de la loi $U[\theta, \theta+1]$ (uniforme sur $[\theta, \theta+1]$) . L'estimateur du maximum de vraisemblance consiste à prendre n'importe quel point entre $[\max X_i - 1, \min X_i]$ ce qui conduit à un bon estimateur, qui converge à vitesse $1/n$, lorsque toutes mes données sont bien tirées suivant une loi $U[\theta, \theta+1]$). En revanche, quand $\theta = 0$ et que l'ajoute une donnée aberrante qui vaut 10, l'intervalle devient vide (la vraisemblance devient nulle) et l'estimateur du maximum de vraisemblance n'existe plus. Une procédure bien plus astucieuse serait de choisir l'estimateur de θ qui maximise

$$L : \theta \mapsto \sum_{i=1}^{n} \mathbb{1}_{[\theta,\theta+1]}(X_i) \tag{1}$$

c'est à dire le nombre d'observations tombant dans l'intervalle $[\theta, \theta+1]$. Notons que si le modèle est exact, les points qui maximisent la vraisemblance sont exactement ceux qui maximisent L. Mais l'avantage du critère L est que même si l'échantillon contient quelques données aberrantes (en fait, même si presque la moitié de l'échantillon est aberrant !), un maximiseur de L continuera à converger à vitesse $1/n$. On n'a donc pas changé la vitesse d'estimation (qui est ici optimale) en prenant ce nouvel estimateur robuste.

Les deux exemples précédents montrent que l'estimateur construit par la méthode du maximum de vraisemblance conduit à des estimateurs qui ne sont pas robustes en général. Il ne s'agit pas d'une spécificité de cette procédure car la plupart des procédures statistiques ne sont pas robustes malheureusement.

Reprenons l'exemple de l'estimation du paramètre θ de la loi uniforme $[0, \theta]$ $\theta > 1/2$ (pour simplifier) dans un contexte bayésien avec la loi a priori de densité

$$\pi_\alpha(\theta) = C\theta^{-\alpha}\mathbb{1}_{(1/2,+\infty)}(\theta) \quad \text{pour un certain } \alpha > 1$$

et une constante de renormalisation $C = C_\alpha > 0$. Un petit calcul montre que la densité a posteriori s'écrit

$$g^L(\theta|\boldsymbol{X}) = (n+\alpha-1)\left((1/2) \vee X_{(n)}\right)^{n+\alpha-1} \theta^{-n-\alpha}\mathbb{1}_{((1/2)\vee X_{(n)},+\infty)}(\theta),$$

et sa fonction de répartition est

$$G^L(\theta|\mathbf{X}) = \left[1 - \left(\frac{(1/2) \vee X_{(n)}}{\theta}\right)^{n+\alpha-1}\right] \mathbb{1}_{((1/2)\vee X_{(n)},+\infty)}(\theta). \tag{2}$$

La loi a posteriori se concentre donc sur les intervalles de la forme

$$\left[(1/2) \vee X_{(n)}, \left(1 + cn^{-1}\right)\left((1/2) \vee X_{(n)}\right)\right] \quad \text{avec } c > 0$$

Si $\theta = 1$ et que l'on ajoute une donnée aberrante qui vaut 100, la loi a posteriori se concentre sur l'intervalle $[100, 100(1 + cn^{-1})]$, ce qui est très mauvais.

Il existe d'autres formes d'instabilité, et pas nécessairement dues à la présence de données aberrantes comme nous allons le voir.

Soit $X_1, \ldots, X_n$ un n-échantillon de loi $P^\star = p^\star \cdot \lambda$ ayant une densité de carré intégrable par rapport à la mesure de Lebesgue λ et prenons comme modèle statistique le modèle $\mathscr{P} = \{P = p \cdot \lambda, p \in \mathcal{D}\}$ où $\mathcal{D}$ est un ensemble de densités sur $\mathbb{R}$ de carré intégrable. Dans ce cas, une méthode d'estimation très classique consiste à chercher la densité $\widehat{p} \in \mathcal{D}$ qui minimise le critère (contraste $\mathbb{L}_2$)

$$p \mapsto G_n(p) = -\frac{2}{n} \sum_{i=1}^{n} p(X_i) + \|p\|^2 .$$

Cette approche qui se justifie par le fait qu'en espérance le critère vaut

$$-\langle p, p^\star \rangle + \|p\|^2 = \|p - p^\star\|^2 - \|p^\star\|$$

et est donc minimal pour $p = p^\star$.

Prenons un exemple très simple où

$$\mathscr{P} = \left\{ P_0 = \mathcal{U}([0,1]), P_1 = \frac{1}{8}\mathcal{U}(0,1/2) + \frac{7}{8}\mathcal{U}(1/2,1) \right\}$$

et l'on prend pour $\mathcal{P}$,

$$p_0 = \mathbb{1}_{[0,1]} \quad \text{et} \quad p_1 : x \mapsto (1/4)\mathbb{1}_{[0,1/2]} + (7/4)\mathbb{1}_{(1/2,1)} + n^2\mathbb{1}_{\{1\}}.$$

Le fait de mettre la valeur $7/4$ en 1 plutôt que n^2 est purement conventionnelle car on ne change pas la loi en modifiant la densité en 1 point.

Supposons que $P^\star = P_0$. Nous obtenons alors que

$$G_n(p_0) = -2 + 1 = -1$$

alors que, en notant N_1, N_2, N_3 le nombre de X_i tombant dans les intervalles $(0, 1/2)$, $(1/2, 1)$ et valant 1 respectivement, nous obtenons que

$$\begin{aligned} G_n(p_1) &= -\frac{2}{n}\left(N_1 \times \frac{1}{4} + N_2 \times \frac{7}{4} + N_3 \times n^2\right) + \|p_1\|^2 \\ &= -\frac{2}{n}\left(N_1 \times \frac{1}{4} + N_2 \times \frac{7}{4} + N_3 \times n^2\right) + \frac{25}{16}. \end{aligned}$$

Comme P_0-p.s., $N_3 = 0$, $N_2 = n - N_1$, on trouve que

$$G_n(p_1) = \frac{3N_1}{n} - \frac{31}{16}$$

et comme sous P_0, $N_n \sim \mathcal{B}(n, 1/2)$, la procédure sélectionne la bonne probabilité parmi le modèle $\mathscr{P}$ avec la probabilité

$$\mathbb{P}\left[G_n(p_1) > G_n(p_0)\right] = \mathbb{P}\left[\frac{3N_n}{n} - \frac{31}{16} > -1\right] = \mathbb{P}\left[\mathcal{B}(n, 1/2) > \frac{5n}{16}\right] \approx 1$$

lorsque n est grand.

Par contre si on ajoute la valeur 1 à l'échantillon,

$$G_n(p_1) \leqslant -\frac{2}{n} N_3 \times n^2 + \frac{25}{16} \leqslant -2n + \frac{25}{16} < -1 = G_n(p_0)$$

dès que $n \geqslant 2$. Il suffit donc d'un point supplémentaire dans mon échantillon pour que la procédure soit instable. En outre, le choix d'une densité particulière pour representer la probabilité P_1 est également source d'instabilité car le pb disparaitrait en choisissant $p_1(1) = 0$ par exemple.

Comment étudier la robustesse d'un estimateur ?

Prenons un modèle statistique parametré $\mathscr{P} = \{P_\theta,\ \theta \in \Theta\}$ pour modéliser la loi $P^\star$ de mes n observations $\boldsymbol{X} = (X_1, \ldots, X_n)$, ce qui signifie que je suppose (à tort ou à raison) que mes observations sont i.i.d. de loi $P^\star = P_{\theta^\star}$ pour un certain $\theta^\star \in \Theta$. Etant donné une fonction de perte ℓ, il est de coutume d'évaluer la performance d'un estimateur $\widehat{\theta}_n$ de $\theta^\star$ par son risque, c'est à dire par la quantité

$$\mathcal{R}(\mathbf{P}^\star, \widehat{\theta}_n) = \mathbb{E}_{\theta^\star}\left[\ell(\widehat{\theta}_n(\boldsymbol{X}), \theta^\star)\right] = \int \ell(\widehat{\theta}_n(\mathbf{x}), \theta^\star) d\mathbf{P}^\star(\mathbf{x})$$

où $\mathbf{P}^\star$ désigne la loi de $\boldsymbol{X}$. En général, les statisticiens étudient ce risque lorsque le modèle est exact, c'est à dire lorsque $\mathbf{P}^\star = P_{\theta^\star}^{\otimes n}$. La question est de savoir ce qu'il devient lorsque :

— $\mathbf{P}^\star = \bigotimes_{i \notin I} P_{\theta^\star} \bigotimes_{\in I} Q_i$ où I est un ensemble de cardinal petit relativement à n et les Q_i sont des lois arbitraires (par exemple des Dirac). Cela permet de remettre en cause l'hypothèse d'équidistribution (mais pas d'indépendance) des données. En outre, cela permet de savoir comment l'estimateur se comporte en la présence d'éventuelles données aberrantes (correspondant à des masses de Dirac).

— $P^\star = (1 - \alpha) P_{\theta^\star} + \alpha Q$ où $\alpha \in (0, 1)$ petit et Q est une loi arbitraire. Les observations restent ici i.i.d. mais on suppose que l'échantillon a été contaminé par un sous-échantillon de loi Q dans une proportion moyenne α. On parle alors de modèle de contamination.

— $\mathbf{P}^\star = \bigotimes_{i=1}^n P_i^\star$ avec $P_i^\star \neq P_{\theta^\star}$ mais $P_i^\star$ proche de $P_{\theta^\star}$ (en un certain sens) pour tout i. Il s'agit là d'étudier la robustesse de l'estimateur à la fois par rapport au modèle $\mathscr{P}$ et à l'hypothèse d'équidistribution.