Des moindres carrés aux moindres déviations

Jean-Jacques Fuchs

IRISA/Université de Rennes I
Campus de Beaulieu – 35042 Rennes Cedex
fuchs@irisa.fr

RÉSUMÉ. *La régression linéaire est un domaine important en pratique qui est, en général, associée aux moindres carrés. Mais on sait depuis longtemps que si les erreurs ne sont pas vraiment gaussiennes et peuvent inclure des valeurs aberrantes il est préférable d'utiliser la norme ℓ_1 et de passer aux moindres déviations. Une version intermédiaire consiste à minimiser la norme ℓ_1 pour les résidus supérieurs à un seuil h et la norme ℓ_2 pour les autres, on retrouve alors la fonction de pénalisation de Huber qui est optimale dans un certain sens. On propose un algorithme qui génère la suite de ces optimums. Le coût considéré dépend d'un paramètre h. L'algorithme démarre en h infini avec l'optimum des moindres carrés qui est simple à obtenir, on propage la solution pour h décroissant, et en h nul, on a l'optimum des moindres déviations.*

ABSTRACT. *Linear regression is mostly dominated by least squares which corresponds to Gaussian noise. But it is known for a long time that if outliers may be present in the measurements, robust regression techniques such as the least absolute deviation method, are preferable. One can also consider an intermediate cost function where residues larger than a threshold h are weighted by the ℓ_1-norm and the others by the ℓ_2-norm. This leads to the Huber penalization that is optimal for a certain contaminated Gaussian distribution. No closed-form solution exist for these cost function and we propose an algorithm which, initialized by the least squares estimate that is optimal for h infinite, builds the sequence of estimates associated with decreasing h, a zero h corresponding the least absolute deviation estimate.*

MOTS-CLÉS : *régression linéaire, moindres carrés, moindres déviations, estimation robuste.*

KEYWORDS: *linear regression, least squares, least deviations, robust estimation*

DOI:10.3166/TS.27.109-119

Extended abstract

Linear regression models are in general associated with the least squares (LS) criterion and Gaussian errors and this leads to a well established theory. But it is known for a long time that if too many measures have large errors, that are not compatible with the Gaussian assumption, one should resort to more robust strategies. The least absolute deviations (LAD) criterion, which is optimal if the errors are double-exponential, is one such strategy. The ℓ_2 norm of the least squares criterion is then replaced by the ℓ_1 norm. The optimum admits no analytical form but can be obtained as the solution of a linear program.

Between the LS and the LAD criteria, there is the so-called Huber function that takes the absolute value of the residues larger than a threshold and the square for the other, smaller, residues. This function, that is theoretically well founded (Huber, 1981), depends thus upon a threshold, say h. It appears as in between the LS criterion, that corresponds to $h = \infty$ and the LAD criterion that corresponds to $h = 0^+$. In this paper, a algorithm is proposed, that minimizes the Huber function and that starting with h infinite and initialized at the LS solution, furnishes all the sequence of optimal regression parameters for decreasing h until the desired h (potentially 0) is attained.

For the regression model

$$b = Ax + e,$$

with A full column rank, one first shows that the optimization problem

$$\min_{x,y} \frac{1}{2}\|Ax - y - b\|_2^2 + h\|y\|_1, \qquad h > 0,$$

corresponds precisely to the minimization of the Huber function with threshold value h, i.e., that the x-part of the optimal solution $\{x,y\}$ is the optimal regression parameter vector associated with the threshold h. The equivalence is known (Geman *et al.*, 1995) and (Fuchs, 1999). This composite criterion can be transformed into a quadratic program and solved using standard routines, but a faster dedicated algorithm, that attains the optimum in a finite number of steps, will be developed.

The algorithm constructs the optimal $\{x,y\}$ for decreasing value of h and stops when the desired value of h is attained. It starts with h infinite and $x = A^+b$ (with $A^+ = (A^TA)^{-1}A^T$) the LS solution and $y = 0$ and propagates this optimal couple for decreasing h. For $h = 0$, the optimal x is then the LAD solution. In between,the optimal x and y vary continuously and are piecewise linear. The real line $]0,\infty[$ is partitioned into a finite number of intervals $]h_k,h_{k+1}]$, and within each interval the optimal x and y vary linearly with h. The algorithm that is proposed is close to the Lasso (Tibshirani, 1996) and associated Least Angular Regression algorithm (Efron *et al.*, 2004) and similar to algorithms proposed in (Rosset *et al.*, 2007).

In a first step, one writes the optimality conditions (8) to be satisfied by $\{x,y\}$. Due to the presence of the ℓ_1-norm in the criterion, the sub-gradient, say u of $\|y\|_1$ at y intervenes in these conditions that are thus difficult to exploit unless one partitions y into its zero components in $\bar{\bar{y}}$ and its non-zero components in $\bar{y}$. This partition of y induces similar partitions on u but also on b and the regression matrix

A, in terms of its rows. With these new notations the optimality conditions in (8) are rewritten as (9) and one can then observe, that if one knows the optimal $\{x, y\}$ for a given h, the relations in (9) allow to extend the optimum to a whole interval in h around the current h. One thus gets expressions of the form (10) for the three quantities that are directly h dependent. The two boundaries of the current interval in h correspond to the values of h for which the current partition of y is no longer valid, this happens either when a component in $\bar{y}$, or more precisely $\bar{y}(h)$ becomes zero, or when a component in $\bar{\bar{u}}(h)$ attains one in absolute value, signaling that the corresponding value of $\bar{\bar{y}}$ which was zero, is about to become non zero.

Eventually to cross such a boundary, one simply has to change the partition of y accordingly, as well as the associated partitions of u, b and A and gets the new expressions (9) and (10) valid in the neighboring interval. One then progresses in this way for decreasing h. For h infinite, the optimum is at $\{x = A^{+}b, y = 0\}$ and one can check that for $h \geqslant h_0 = \|(AA^{+} - I)b\|_{\infty}$, the expressions in (9) are admissible with $\bar{\bar{b}} = b$, $\bar{\bar{A}} = A$ and the corresponding $\bar{\bar{u}} = u$. This value of h is thus the upper boundary of the first interval. One can add that for decreasing h, the number of rows in $\bar{\bar{A}}$ generically decreases, but may augment locally and that for h small and close to zero, in the last interval, $\bar{\bar{A}}$ is generically square and invertible.

As a conclusion, one may add that disposing of the whole set of solutions is both interesting and dangerous. It is interesting because it allows to adapt the threshold h *a posteriori* and thus get a robust solution, dangerous because this very possibility may influence the decision maker and thus introduce some prejudice into the solution.

1. Introduction

Les techniques de régression linéaire ont longtemps été dominées par les moindres carrés (MC) car ils sont simples à mettre en oeuvre et basés sur une théorie bien établie. L'approche des MC est optimale si les erreurs sont gaussiennes. Mais on sait depuis longtemps que les estimées obtenues par les MC sont rapidement sans signification si certaines des mesures sont aberrantes, si elles présentent des erreurs importantes en proportion plus grande que ne le permet l'hypothèse gaussienne. On utilise alors des approches robustes qui atténuent l'influence de ces données aberrantes. Parmi ces méthodes celle des moindres déviations (MD), dans laquelle la norme euclidienne (ℓ_2) est remplacée par la norme ℓ_1, est la plus simple à mettre en œuvre, car elle ne nécessite pas de réglage, elle ne fait pas intervenir de seuil, par exemple.

C'est notamment Laplace, qui a étudié (1793) l'approche des MD, qui est optimale si les erreurs suivent la loi de Laplace et cela s'est donc passé avant l'étude des MC réalisée par Legendre (1805) et Gauss (1823). Contrairement aux MC, l'optimum des MD ne peut être obtenu qu'à l'aide d'un algorithme. De nombreux algorithmes ont été proposés. On peut par exemple citer les algorithmes de programmation linéaire ou des variantes (Barrodale *et al.*, 1973 ; Li *et al.*, 2004 ; Bloomfield *et al.*, 1983 ; Rusinsky *et al.*, 1989).

L'absence de paramètre ou de seuil à régler peut être aussi perçu comme un désavantage et entre le critère des MC (la somme des carrés des écarts) et celui des MD (la somme des valeurs absolues des écarts), on trouve la fonction de Huber qui consiste à prendre la valeur absolue pour les écarts supérieurs à un seuil et le carré pour les autres. Quand ce seuil va de plus l'infini à zéro, ce critère passe de façon continue de celui des MC à celui des MD. La fonction de Huber a aussi une justification théorique, elle correspond à maximiser la log-vraisemblance pour des erreurs centrées et indépendantes dont la densité de probabilité est de la forme, $p(e) = (1-\epsilon)\zeta(e) + \epsilon q(e)$ avec $\zeta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ la Gaussienne standard, et $q(e)$ une densité qui reste à déterminer. Il s'agit donc d'une densité du type gaussienne contaminée, où la densité q est choisie de façon à minimiser l'information (de Fisher) contenue dans p. Ce problème d'optimisation fonctionnelle a une solution analytique (Huber, 1981) et la densité p résultante peut s'écrire sous la forme :

$$\begin{aligned} p(e) &= \frac{1-\epsilon}{\sqrt{2\pi}} \exp(-e^2/2) \qquad |e| \leqslant h \\ &= \frac{1-\epsilon}{\sqrt{2\pi}} \exp(h^2/2 - h|e|) \qquad |e| > h \end{aligned} \qquad [1]$$

où le seuil h dépend de ϵ, le taux de contamination, et varie de 0 à ∞ quand ϵ va de 1 à 0. Le seuil h est tel que la densité $p(e)$ est bien de mesure 1 et il vérifie donc

$$\int_{-\infty}^{\infty} p(x)dx = (1-\epsilon) \int_{-h}^{h} \zeta(x)dx + 2\frac{1-\epsilon}{h}\zeta(h) = 1.$$

En prenant le logarithme de cette densité, on a bien un critère de la forme annoncée, quadratique pour $|e| < h$ et linéaire ensuite, le critère et sa dérivée étant de plus continus en $e = |h|$. Dans la suite, nous proposons un algorithme qui permet d'optimiser ce critère et qui, partant de h infini et initialisé à l'optimum des MC, fournit la suite des optimums quand h décroît vers zéro.

2. Généralités

On considère le modèle de régression linéaire suivant

$$b = Ax + e \qquad [2]$$

où b est le vecteur de dimension m des observations, x le vecteur de dimension n à estimer, et où A est la matrice des régresseurs de dimension (m,n) et de rang colonne plein. On peut voir e comme représentant les erreurs de mesure qui font que b n'appartient pas à l'image de A, n'est pas égal à Ax.

Si on peut supposer que ces erreurs sont des échantillons indépendants d'une densité gaussienne de moyenne nulle et de même variance, l'optimum au sens du maximum de vraisemblance (MV) consiste à minimiser à l'aide de x la quantité $\sum r_i^2$ avec r_i la i-ème composante du vecteur des résidus $r = Ax - b$ et on obtient l'estimée au sens des MC. Si on considère que les erreurs suivent une loi de Laplace, l'optimum au sens du MV consiste à minimiser $\sum |r_i|$ et on obtient l'estimée au sens

des MD. Et, enfin, si les erreurs ont pour densité (1), l'optimum consiste à minimiser $\sum f(r_i)$ où la fonction $f(.)$ est de la forme

$$f(r_i) = \frac{r_i^2}{2} \boldsymbol{I}_{|r_i| \leqslant h} + (h|r_i| - \frac{h^2}{2}) \boldsymbol{I}_{|r_i| > h} \quad [3]$$

où $\boldsymbol{I}$ désigne la fonction indicatrice. On peut noter que cette fonction $f(.)$ est continue et à dérivée première continue.

Dans la suite, on va s'intéresser au problème d'optimisation suivant :

$$\min_{x,y} \frac{1}{2} \|Ax - y - b\|_2^2 + h\|y\|_1, h > 0, \quad [4]$$

dont on va montrer qu'il est équivalent à la minimisation du critère d'Huber appliqué à (2). Il s'agit donc d'établir que (4) est bien équivalent à :

$$\min_x \sum_i f(r_i) \qquad \text{avec} \quad r = Ax - b. \quad [5]$$

On remarque d'abord que (4) est séparable, que l'on peut commencer par optimiser par rapport aux composantes de y. En effet, le problème (4) s'écrit aussi :

$$\min_{x,y} \sum_i \frac{1}{2}(r_i - y_i)^2 + h|y_i|, r = Ax - b. \quad [6]$$

Comme y_i est seulement présent dans le i-ème terme de la somme, on utilise y_i pour minimiser ce i-ème terme. Un petit exercice non trivial permet alors de trouver que l'optimum est atteint en $y_i = 0$ si $h > |r_i|$ et que sinon il est atteint en $y_i = r_i - h$ signe(r_i). Le minimum du i-ème terme vaut par conséquent $r_i^2/2$ si $|r_i| \leqslant h$ et $h|r_i| - h^2/2$ si $|r_i| \geqslant h$ ce qui transforme exactement (4) en (5) ou il reste à prendre le minimum en x. Cette équivalence est connue, voir par exemple (Geman *et al.,* 1999) où (6) est déduit de (5) en utilisant des outils d'analyse convexe. Dans (Fuchs, 1999), on introduit les variables y dans (4) comme une façon de modéliser et localiser les erreurs aberrantes dans un contexte de régressions linéaires (2) et on constate ensuite l'équivalence entre (4) et le critère de Huber.

Pour résoudre le problème d'optimisation (4), on peut observer que l'on peut se débarrasser de $\|y\|_1 = \sum |y_i|$, en doublant le nombre d'inconnues y_i, en introduisant $y_i^+ = \max(y_i, 0)$ et $y_i^- = \max(-x_i, 0)$. On a alors $|y_i| = y_i^+ + y_i^-$ et $y_i = y_i^+ - y_i^-$. En faisant ces substitutions, le terme de pénalité devient linéaire et le critère devient quadratique mais il faut lui adjoindre les contraintes $y_i^+ \geqslant 0$ et $y_i^- \geqslant 0$. On obtient un programme quadratique (Fletcher, 1987) que l'on sait résoudre avec des algorithmes efficaces et robustes.

L'objet de la suite est de développer un algorithme dédié bien plus rapide qui minimise (4) en un nombre fini d'étapes. Il construit l'optimum de (4) pour des valeurs décroissantes de h et s'arrête quand le h souhaité est atteint. Il démarre en h infini par la solution des MC pour x et $y = 0$ et propage cet optimum pour h décroissant. En $h = 0$, on sait déjà que l'optimum en x est la solution des MD. Entre les deux, on va voir que x et y varient de façon continue et linéaire par morceaux, on va partitionner $]0, \infty[$ en un nombre fini d'intervalles et, à l'intérieur de chacun d'eux, x et y varient de façon linéaire.

On peut aussi retrouver intuitivement les optimums de (4) aux deux extrémités. Pour h infini, y ne peut être que nul à l'optimum et x minimise le critère restant. Quand h tend vers zéro, y est presque gratuit et on vise un coût global quasiment nul car on dispose de $n + m$ degrés de liberté. Le premier terme du critère est alors nul à l'optimum, ce qui donne $Ax - y - b = 0$ ou aussi $y = Ax - b$ et on minimise le second $\|y\|_1$ sous $y = Ax - b$, ce qui donne, pour x, la solution des MD.

3. Algorithme d'optimisation

3.1. Introduction

On développe un algorithme qui résout (4) et donc (5) en un nombre fini de pas. De nombreux algorithmes permettent de minimiser (5) , on a déjà cité les algorithmes de programmation quadratique mais on a également des algorithmes du type moindres carrés re-pondérés itérés (IRLS) (Holland *et al.*, 1977 ; Rusinsky *et al.*, 1989) ou d'autres méthodes itératives (Geman *et al.*, 1995). Une version similaire à l'algorithme développé ici, est par ailleurs déductible des algorithmes proposés dans (Rosset *et al.*, 2007).

On va établir que l'optimum $\{x, y\}$ de (4) ou plus précisément $\{x(h), y(h)\}$ est une fonction continue et linéaire par morceaux de h. On va décomposer l'axe des réels en un nombre fini (de l'ordre de n) intervalles (h_{k+1}, h_k) telle que à l'intérieur de chaque intervalle on ait par exemple pour $x(h)$ une expression de la forme $x(h) = X_1 + hX_2$ avec X_i des vecteurs constants.

On va en fait construire $x(h)$ pour des valeurs décroissantes de h en commençant par $h \geqslant h_0$ où h_0 reste à définir et pour lequel l'optimum est constant et atteint en la solution bien connue des MC : $x(h) = (A^T A)^{-1} A^T b$ et $y(h) = 0$, et, en progressant, on définit les bornes des intervalles en h et les valeurs de $x(h)$ valides dans ces intervalles. Dans un premier temps, on obtient une expression de X_1 et X_2 dans $x(h) = X_1 + hX_2$ valide pour h légèrement inférieure à h_0, puis on définit la valeur de $h_1 < h_0$ la borne inférieure de l'intervalle pour laquelle cette expression cesse d'être valide, puis la nouvelle expression de $x(h)$ et ainsi de suite...

3.2. Conditions d'optimalité

Le problème d'optimisation (4) est convexe, les conditions d'optimalités du premier ordre sont à la fois nécessaires et suffisantes. Comme il n'y a pas de contrainte, on écrit que le gradient par rapport à x et le sous-gradient par rapport à y sont nuls. On a alors :

$$A^T (Ax - y - b) = 0 \qquad \text{et} \qquad Ax - y - b - hu = 0,$$

avec u, un sous-gradient de $\|y\|_1$ en y, un vecteur de la dimension de y qui satisfait (Fletcher, 1987) :

$$u_i = \text{signe}(y_i) \qquad \text{si} \qquad y_i \neq 0 \qquad \text{et} \qquad |u_i| \leqslant 1 \text{ sinon.} \qquad [7]$$

Pour $h > 0$, les conditions d'optimalité peuvent se simplifier en

$$Ax - y - b - hu = 0 \qquad \text{et} \qquad A^T u = 0. \qquad [8]$$

Les relations (8) sont difficiles à exploiter à cause de la présence de u qui n'est pas défini de façon unique pour les composantes nulles de y. On partitionne donc y en deux sous-vecteurs, $\bar{y}$ dont les composantes sont non nulles et $\bar{\bar{y}}$ dont les composantes sont nulles. On applique cette même partition à u et on a alors $\bar{u} = \text{signe}(\bar{y})$ et $\|\bar{\bar{u}}\|_\infty \leqslant 1$, voir (7), où $\|u\|_\infty = \max_j |u_j|$ désigne la norme ℓ_∞ de u. On étend aussi cette partition à b et à la matrice A, à ses lignes. On note ainsi, par exemple $\bar{A}$ la sélection des lignes de A associées aux composantes de $\bar{y}$.

Il faut noter que c'est la partition de y qui pilote toutes les autres. Avec ces notations, les $n + m$ relations dans (8) deviennent

$$\begin{aligned} &\bar{A}x - \bar{y} - \bar{b} - h\bar{u} = 0, \\ &\bar{\bar{A}}x - \bar{\bar{b}} - h\bar{\bar{u}} = 0, \\ &\bar{A}^T\bar{u} + \bar{\bar{A}}^T\bar{\bar{u}} = 0. \end{aligned} \qquad [9]$$

Ces équations sont maintenant exploitables comme nous allons le voir plus loin. Mais on a bien sûr supposé connaître la partition de y associée à la valeur de h considérée.

3.3. Développement

Si on suppose connaître l'optimum de (4) pour une certaine valeur de h, les relations (9) permettent d'étendre cet optimum au voisinage et de trouver les bornes de l'intervalle en h dans lequel cette extension est valide. On peut alors franchir ces frontières et trouver l'expression de l'optimum dans les intervalles voisins. Il reste finalement à savoir initialiser cette procédure pour développer l'algorithme qui permet de résoudre (4) pour tout h. Comme nous l'avons déjà indiqué, cette initialisation ne pose pas de problème car pour h grand, (4) se réduit au problème des moindres carrés.

L'optimum de (4) pour un h donné, est un couple $\{x, y\}$. On peut lui adjoindre le vecteur sous-gradient u déduit de la première relation de (8), par exemple et qui, bien sûr, satisfait alors (7).

On peut alors décomposer ce triplet optimal $\{x, y, u\}$ redondant en $\{x, \bar{y}, \bar{\bar{u}}\}$ et $\{\bar{\bar{y}} = 0, \bar{u} = \text{sign } \bar{y}\}$ où le premier ensemble de dimension $n+m$ (comme le couple optimal) contient toute l'information qu'il s'agit d'étendre au voisinage et le second est précisément constitué des variables qui restent invariantes dans le voisinage. Les relations (9), qui traduisent les conditions nécessaires et suffisantes, forment alors un système de $n+m$ équations linéaires en $n+m$ inconnues $\{x, \bar{y}, \bar{\bar{u}}\}$, dépendant de h. On peut récrire le système (9) sous la forme échelonnée suivante :

$$\begin{aligned} &\bar{\bar{A}}^T\bar{\bar{A}}x = \bar{\bar{A}}^T\bar{\bar{b}} - h\bar{A}^T\bar{u} \\ &\bar{A}x - \bar{y} = \bar{b} + h\bar{u} \\ &\bar{\bar{A}}x - h\bar{\bar{u}} = \bar{\bar{b}} \end{aligned}$$

où la première équation ne dépend que x, la seconde de x et $\bar{y}$. Si $\bar{\bar{A}}$ est de rang colonne plein, ce que l'on va supposer et qui est génériquement vrai, le système a une

solution unique de la forme

$$\begin{aligned} x(h) &= X_1 + hX_2 \\ \bar{y}(h) &= V_1 + hV_2 \\ h\bar{\bar{u}}(h) &= W_1 + hW_2 \end{aligned} \qquad [10]$$

où $V_.$, $W_.$ and $X_.$ sont des vecteurs constants de dimensions adéquates, que nous ne détaillons pas tous, on a par exemple

$$X_1 = \bar{\bar{A}}^{+}\bar{\bar{b}} \qquad \text{et} \qquad X_2 = (\bar{\bar{A}}^T\bar{\bar{A}})^{-1}\bar{A}^T\bar{u}.$$

Ces relations décrivent comment le triplet $\{x, \bar{y}, \bar{\bar{u}}\}$ évoluent en fonction de h. Elles sont valides aussi longtemps que la partition induite par y reste valide, aussi longtemps que le second jeu de paramètres $\{\bar{\bar{y}}, \bar{u}\}$ reste lui aussi valide.

La seconde relation dans (10) dit comment les composantes non nulles de y évoluent quand h varie autour de la valeur courante. Cette relation cesse d'être valide dès qu'une composante de $\bar{y}$ devient nulle. De la même façon, la dernière relation est valide aussi longtemps qu'aucune composante de $\bar{\bar{u}}$ n'atteint un en valeur absolue. En effet, voir (7), si une composante de $\bar{\bar{u}}$ devient égale à un, par exemple, cela signifie que la composante correspondante dans $\bar{\bar{y}}$, qui est égale à zéro va devenir positive.

Quand h décroît (ou croît) à partir de sa valeur courante, il faut donc détecter lequel des deux événements arrive en premier : une composante de $\bar{y}$ qui s'annule ou une composante de $\bar{\bar{u}}$ qui atteint un, en valeur absolue, la valeur de h associée, sera alors la borne inférieure (supérieure) de l'intervalle de validité des relations dans (10).

En une telle borne la partition de y est modifiée et il faut changer toutes les autres partition en conséquence. Pour A par exemple, dans le premier cas, on enlève la ligne de $\bar{A}$ associée à la composante de $\bar{y}$ devenant nulle et on la rajoute dans $\bar{\bar{A}}$, dans le second cas une ligne de $\bar{\bar{A}}$ est déplacée vers $\bar{A}$. On verra que pour h grand, $\bar{\bar{A}} = A$ de rang colonne plein par hypothèse. Quand h décroît, le nombre de ligne dans $\bar{\bar{A}}$ diminue en général, mais peut aussi augmenter localement. Et, pour $h > 0$ mais suffisamment petit, $\bar{\bar{A}}$ est, génériquement, une matrice carrée inversible.

Si on résout alors le nouveau système linéaire échelonné, on trouve les nouvelles versions des relations (10) qui sont valides dans l'intervalle voisin puisque les valeurs de $\{x, \bar{y}, \bar{\bar{u}}\}$ ainsi obtenues et le nouveau couple $\{\bar{\bar{y}} = 0, \bar{u} = \text{sign } \bar{u}\}$ satisfont par construction les conditions nécessaires et suffisantes (9).

On peut noter que pour les valeurs de h correspondant aux bornes de ces intervalles, le x et y optimales admettent deux expressions donnant la même valeur. On a donc bien un optimum $x(h)$ qui est continu et linéaire par morceaux.

3.4. Initialisation pour h grand

Nous avons déjà vu que, pour$h$ infini, l'optimum de (4) est en $y = 0$ et $x = A^{+}b$ où $A^{+} = (A^TA)^{-1}A^T$ la pseudo-inverse de A, la solution du problème des MC $\min_x \|Ax - b\|_2^2$.

Vérifions que le couple $\{x = A^{+}b, y = 0\}$ satisfait (9) pour h suffisamment grand et cherchons la borne inférieure h_0 de cet intervalle. Dans (9), le premier jeu d'équations disparaît, le second avec $\bar{\bar{A}} = A$ donne la valeur de $\bar{\bar{u}}$ en fonction de h

$$\bar{\bar{u}}(h) = \frac{1}{h}(\bar{\bar{A}}\bar{\bar{A}}^{+} - I)b,$$

et cette valeur satisfait le troisième jeu d'équations. La valeur de $\bar{\bar{u}}(h)$ ainsi obtenue satisfait (7) si

$$h \geqslant h_0 = \|(\bar{\bar{A}}\bar{\bar{A}}^{+} - I)b\|_{\infty} = \|(AA^{+} - I)b\|_{\infty},$$

qui est donc la borne inférieure h_0 recherchée.

Pour aller au-delà cette borne et passer à l'étape standard décrite plus haut, il suffit de définir la nouvelle partition de y valide dans l'intervalle $[h_1, h_0]$ suivant, où h_1 n'est pas encore connu. Soit $j_1 = \arg\max |r_j|$ avec $r = (AA^{+} - I)b$, l'indice de la composante de y qui va devenir non nulle en h_0^{-}, on enlève alors la composante j_1 de $\bar{\bar{y}}$ dont la dimension passe à m-1 pour créer $\bar{y} = 0$ de dimension 1 et de la même façon on enlève à $\bar{\bar{u}}$ sa j-ème composante pour créer $\bar{u} = \text{sign}(r_{j_1})$. Les autres quantités $\bar{b}$, $\bar{\bar{b}}$, $\bar{A}$ et $\bar{\bar{A}}$ suivent sans difficulté.

3.5. Quand h décroît vers zéro

Quand h tend vers zéro, on a déjà vu que l'optimum de (4) tend vers l'optimum du problème des moindres déviations. On peut notamment déduire ce résultat de l'équivalence de (4) et de (5), mais on peut également l'établir directement.

Les conditions d'optimalité du problème des moindres déviations

$$\min_{x} \|Ax - b\|_1, \qquad [11]$$

sont

$$A^{T}u = 0, Ax - b = y,$$

avec u un sous-gradient de $\|y\|_1$ en y, satisfaisant (7). Et on constate donc bien que le triplet $\{x, y, u\}$ ainsi obtenu est bien une solution de (8) pour $h = 0$.

Pour compléter ce point, remarquons que (11) peut se mettre sous la forme d'un programme linéaire et que l'on peut alors déduire de la théorie associée, que l'optimum est génériquement atteint pour un point x solution exacte d'un sous-ensemble de n équations linéaires indépendantes extraites des m équations présentes dans $Ax = b$. Cela signifie que y sera identiquement nul en ces mêmes n composantes. Avec les notations introduites, on a alors l'optimum $x(0) = \bar{\bar{A}}^{-1}\bar{\bar{b}}$ et par conséquent $\bar{y}(0) = \bar{A}\bar{\bar{A}}^{-1}\bar{\bar{b}} - \bar{b}$.

Si on utilise l'algorithme proposé pour résoudre le problème des MD, on a alors, dans l'ultime intervalle en h de borne inférieure égale à 0, les relations suivantes (10)

$$x(h) = \bar{\bar{A}}^{-1}\bar{\bar{b}} + h(\bar{\bar{A}}^T\bar{\bar{A}})^{-1}\bar{A}^T\bar{u}$$
$$\bar{y}(h) = \bar{A}\bar{\bar{A}}^{-1}\bar{\bar{b}} - \bar{b} + h\bar{A}(\bar{\bar{A}}^T\bar{\bar{A}})^{-1}\bar{A}^T\bar{u} - h\bar{u}$$
$$\bar{\bar{u}}(h) = -\bar{\bar{A}}^{-T}\bar{A}^T\bar{u}.$$

4. Conclusions

En absence d'hypothèse sur l'erreur de mesure, on utilise en général le critère des moindres carrés pour résoudre une régression linéaire multiple. Mais en présence de mesures aberrantes, il peut être intéressant de choisir une fonction à croissance moins rapide que le carré pour les résidus supérieurs à un seuil. On a proposé un critère (4), fonction d'un paramètre h qui, quand h va de plus l'infini à zéro, passe du critère des moindres carrés (pour h grand) à celui des moindres déviations (pour h nul) en passant par le critère de Huber pour les valeurs intermédiaires. Et, on a développé un algorithme qui fournit l'ensemble des solutions, fonction de h, de toute cette gamme de problème. Il s'agit d'un algorithme facile à mettre en œuvre qui donne la solution exacte et qui, en fait, décompose l'axe réel positif en sous-intervalles dans lesquels l'optimum est linéaire en h. Disposer de l'ensemble des solutions est intéressant, car cela permet d'adapter le seuil *a posteriori* et d'obtenir une solution robuste.

Mais c'est également une option à manipuler avec précaution si on ne veut pas influencer le résultat avec des *a priori* forcément subjectifs.

Bibliographie

Barrodale I., Roberts F. (1973). « An improved algorithm for Discrete ℓ_1 linear Approximation », *SIAM J. Num. Analysis*, vol. 10, n° 5, p. 839-848.

Bloomfield P., Steiger W. (1983). *Least Absolute Deviations : Theory, Applications and Algorithms*, Birkauser, Boston.

Efron B., Johnstone I., Hastie T., Tibshirani R. (2004). « Least angle regression », *Annals of Statistics*, vol. 32, p. 407-499.

Fletcher R. (1987). *Practical Methods of Optimization*, John Wiley & Sons, New York.

Fuchs J.-J. (1999). « A new approach to robust linear regression », *14th IFAC World Congress*, Beijing, p. 427-432, july.

Geman D., Yang C. (1995). « Nonlinear image recovery with half-quadratic regularization », *IEEE Transactions on Image Processing*, vol. 7, n° 7, p. 932-946.

Holland P., Welsh R. (1977). « Robust regression using iteratively reweighted least squares », *Comm. Statistics*, vol. 6, n° 9, p. 813-828.

Huber P. (1981). *Robust statistics*, 2nd edn, John Wiley & Sons.

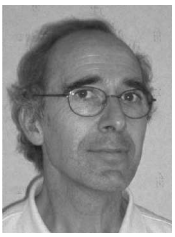
Li Y., Arce G. (2004). « A Maximum Likelihood Approach to Least Absolute Deviation Regression », *J. of Appl. Sign. Proc.*, vol. 2004, n° 12, p. 1762-1769.

Rosset S., Zhu J. (2007). « Piecewise Linear Regularized Solution Paths », *The Annals of Statistics*, vol. 35, n° 3, p. 1012-1030.

Rusinsky S., Olsen E. (1989). « L_1 and L_∞ Minimization via a Variant of Karmarkar's algorithm », *IEEE Transactions on Signal Processing*, vol. 37, n° 2, p. 245-253.

Tibshirani R. (1996). « Regression shrinkage and selection via the lasso. », *J. Royal. Statist. Soc. B.*, vol. 1, p. 267-288.

Reçu 1/10/2009
Accepté 15/05/2010



Jean Jacques Fuchs. *Ingénieur Supelec (1973) et Master du MIT (1974) a d'abord rejoint Thomson-CSF puis assez rapidement l'IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires) en 1976. Depuis 1983, il est professeur à l'Université de Rennes 1. Ses thèmes de recherche sont passés de l'automatique, l'identification et la commande adaptative, dans lequel il a soutenu une thèse d'État en 1982, au traitement du signal où il s'intéresse plus particulièrement au traitement d'antenne, à la détection et aux représentations parcimonieuses.*