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.

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 supposées 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 oeuvre, 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 et de nombreux algorithmes ont été proposés. On peut par exemple utiliser les algorithmes de programmation linéaire ou des variantes [1, 2, 3, 4].

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 [5] et la densité p résultante peut s'écrire sous la forme :

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

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

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

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 continues en $e = |h|$. Dans la suite, nous proposons un algorithme 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

$$b = Ax + e \tag{2}$$

où b est le vecteur de dimension m des observations, x le vecteur de dimension n à estimer, et la matrice A des régresseurs est de dimension (m, n) et de rang colonne plein. On peut voir e comme représentant les erreurs de mesure qui font que b 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 $\sum r_i^2$ 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

$$f(r_i) = \frac{r_i^2}{2} \mathbb{1}_{|r_i| \leq h} + (h|r_i| - \frac{h^2}{2}) \mathbb{1}_{|r_i| > h} \qquad (3)$$

où $\mathbb{1}$ désigne la fonction indicatrice. On peut noter que cette fonction $f(.)$ est continue et à dérivée première continue.

Dans la suite nous allons nous intéresser au problème d'optimisation suivant :

$$\min_{x,y} \frac{1}{2} \|Ax - y - b\|_2^2 + h\|y\|_1, \qquad h > 0, \qquad (4)$$

dont nous allons 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 :

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

On remarque d'abord que (4) est séparable et peut s'écrire :

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

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 $y_i^* = r_i - h$ sign r_i. Le minimum du i-ème terme vaut par conséquent $r_i^2/2$ si $|r_i| \leq h$ et $h|r_i| - h^2/2$ si $|r_i| \geq h$ ce qui transforme exactement (4) en (5) ou il reste à prendre le minimum en x. Cette équivalence est connue, voir par exemple [6] où (6) est déduit de (5) en utilisant des outils sophistiqués d'analyse convexe. Nous avons introduit les variables y dans (4) comme une façon de modéliser et localiser les erreurs aberrantes [7] dans un contexte des régressions linéaires (2) et n'avons réalisé que plus tard l'équivalence entre (4) et le critère de Huber.

Nous observons maintenant que (4) est un problème convexe qui peut se mettre sous la forme d'un programme quadratique. Dans la suite, nous développons un algorithme dédié qui minimise (4) en un nombre fini d'étapes. Il suit l'optimum de (4) pour des valeurs décroissantes de h et on l'arrête quand le h souhaité est atteint. Intuitivement, pour h infini, le second terme du critère est nul à l'optimum, ce qui donne $y = 0$ et on utilise x pour minimiser le premier terme $\|Ax - b\|_2^2$ ce qui donne la solution des MC. De façon analogue, quand h tend vers zéro, le premier terme dans le critère est nul à l'optimum, ce qui donne $Ax - y - b = 0$ ou aussi $y = Ax - b$ et on minimise le second $\|y\|_1 = \|Ax - b\|_1$ ce qui donne, pour x, la solution des MD. Et, entre les deux, nous allons 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.

3 Algorithme d'optimisation

3.1 Introduction

Nous développons un algorithme qui résout (4) et donc (5) en un nombre fini de pas. De nombreux algorithmes permettent de minimiser (5) , on peut citer des algorithmes du type moindres carrés re-pondérés itérés (IRLS) [8, 4] ou d'autres méthodes itératives [6]. Une version similaire à l'algorithme que nous développons, est par ailleurs déductible des algorithmes proposés dans [9].

Nous allons montrer que l'optimum $\{x, y\}$ de (4) ou plus précisément $\{x(h), y(h)\}$ est une fonction continue et linéaire par morceaux de h. Nous allons 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 $h \geq h_0$ où h_0 reste à définir et pour lequel l'optimum est constant et atteint en $x(h) = (A^T A)^{-1} A^T b$ et $y(h) = 0$, et en progressant nous définirons les intervalles en h et les valeurs de $x(h)$ valides dans ces intervalles. Dans un premier temps nous allons donc obtenir une expression de X_1 et X_2 dans $x(h) = X_1 + hX_2$ valide pour h légèrement inférieure à h_0, puis nous allons définir 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 :

$$A^T(Ax - y - b) = 0 \quad \text{et} \quad Ax - y - b - hu = 0, \qquad (7)$$

avec u, un sous-gradient de $\|y\|_1$ en y, un vecteur de la dimension de y qui satisfait [10] :

$$u_i = \text{sign}(y_i) \text{ si } y_i \neq 0 \quad \text{et} \quad |u_i| \leq 1 \text{ sinon} \qquad (8)$$

Les relations (7) 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. Nous partitionnons donc y en ses composantes non nulles dans $\bar{y}$ et nulles dans $\bar{\bar{y}}$. Cela induit une partition de u, en $\bar{u}$=sign$(\bar{y})$ et $\|\bar{\bar{u}}\|_\infty \leq 1$, voir (8). On peut alors remplacer y dans (7) par $\bar{I}\bar{y}$ où $\bar{I}$ est la sélection de colonnes de la matrice identité I telle que $y = Iy = \bar{I}\bar{y}$. On introduit aussi la notation $\bar{A} = \bar{I}^T A$ et $\bar{\bar{I}}$ et $\bar{\bar{A}}$ associé à $\bar{\bar{y}}$. Il faut noter que c'est la partition de y qui pilote toutes les autres. Avec ces notations, les n+m relations dans (7) deviennent

$$A^T(Ax - \bar{I}\bar{y} - b) = 0$$

$$\begin{aligned} \bar{A}\, x - \bar{y} - \bar{b} - h\bar{u} &= 0, \\ \bar{\bar{A}} x - \bar{\bar{b}} - h\bar{\bar{u}} &= 0. \end{aligned} \tag{9}$$

Elles sont maintenant parfaitement exploitables comme nous allons le voir plus loin. Mais on a bien sur supposé que l'on connaissait la partition de y. Pour $h > 0$, on peut remplacer la première relation dans (9) par $A^T u = 0$, qui se déduit facilement de (7) et qui s'écrit plus précisément :

$$\bar{A}^T \bar{u} + \bar{\bar{A}}^T \bar{\bar{u}} = 0.$$

3.3 Développement

Si on suppose maintenant connaître l'optimum de (4) pour une certaine valeur de h, les relations (9) permettent d'étendre cet optimum au voisinage et même de trouver les bornes de l'intervalle en h dans lequel cette extension est justifiée. 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 seconde relation de (7), par exemple et qui, bien sûr, satisfait alors (8).

On peut alors décomposer ce triplet optimal $\{x,\ y,\ u\}$ redondant en $\{x,\ \bar{y},\ \bar{\bar{u}}\}$ et $\{\bar{\bar{y}} = 0,\ \bar{u} = \text{sign } \bar{\bar{y}}\}$ 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 3 relations suivantes, qui traduisent les conditions nécessaires et suffisantes,

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

forment alors un système de n+m équations linéaires en n+m inconnues $\{x,\ \bar{y},\ \bar{\bar{u}}\}$ dont le second membre est connu mais dépend de h. On peut récrire ce système 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 $\bar{y}$ et ainsi de suite. Si $\bar{\bar{A}}$ est de rang colonne plein, il a donc toujours une solution unique, qui est de la forme

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

où $V_\cdot, W_\cdot$ and $X_\cdot$ sont des vecteurs constants de dimensions adéquates, que nous ne détaillons pas tous, on a par exemple $X_1 = \bar{\bar{A}}^+ \bar{\bar{b}}$ et $X_2 = (\bar{\bar{A}}^T \bar{\bar{A}})^{-1} \bar{A}^T \bar{u}$. 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 (8), dès qu'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 surveiller 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 supérieure (inférieure) de l'intervalle de validité des relations dans (10).

Pour passer une telle borne il faut changer toutes les partitions, 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 $\bar{A}$. Il faut évidemment faire les modifications associées dans $\bar{u}$ et $\bar{\bar{u}}$ et dans $\bar{y}$ et $\bar{\bar{y}}$, dans u on déplace une composante qui vaut ± 1 et dans y on déplace une composante qui vaut 0.

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 bien les conditions nécessaires et suffisantes (7).

On peut noter que pour les valeurs de h correspondant à des 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 indiqué que -intuitivement- pour h infini, l'optimum est en $y = 0$ et (4) se réduit à $\min_x \|Ax - b\|_2^2$ le problème des moindres carrés avec pour optimum $x = A^+ b$ où $A^+ = (A^T A)^{-1} A^T$ la pseudo-inverse de A supposé de rang colonne plein. Pour trouver la borne inférieure h_0 de cet intervalle, il suffit de résoudre (7) avec $y = 0$, on obtient $x = A^+ b$ et $hu = (AA^+ - I)b$. Ce triplet $\{x,\ y,\ u\}$ est une solution valide de (7) et donc bien l'optimum de (4), pourvu que le $u = \bar{\bar{u}}$ ainsi défini vérifie $\|u\|_\infty \leq 1$. En effet, à $y = \bar{\bar{y}} = 0$ on doit associer $u = \bar{\bar{u}}$ un vecteur dont toutes les composantes doivent être inférieures ou égales à un en valeur absolue, voir (8). Puisque $h\|u\|_\infty = \|r\|_\infty$ avec $r = (AA^+ - I)b$, c'est le cas aussi longtemps que $h \geq h_0 = \|r\|_\infty$.

Nous avons donc trouver h_0 la borne inférieure de l'intervalle et pour aller au delà de 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.

Soit $j_1 = \arg\max |r_j|$ avec r_j la j-ème composante de r. On enlève alors la composante j_1 de $\bar{\bar{y}}$ dont la dimension passe à n-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

Nous avons déjà indiqué que quand h tend vers zéro l'optimum de (4) tend vers l'optimum du problème des moindres déviations, $\min_x \|Ax - b\|_1$. Pour démontrer ce résultat, on récrit ce dernier sous la forme

$$\min_{x,y} \|y\|_1 \quad \text{sous} \quad y = Ax - b. \tag{11}$$

les conditions d'optimalité de ce problème sont, en introduisant le lagrangien et après quelques manipulations, $A^T u = 0,\ Ax - b = y$ avec u un sous-gradient de $\|y\|_1$ en y, satisfaisant (8). Et on constate donc bien que le triplet $\{x,\ y,\ u\}$ ainsi obtenu est une solution valide de (7) pour $h \to 0$. On rappelle que (7) peut se récrire $hA^T u = 0, Ax - b - y = hu$.

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 atteint pour un x solution exacte d'un sous-ensemble de n équations linéaires extraites des m équations présentes dans $Ax = b$. Avec nos notations, cela signifie que le x optimal satisfait $\bar{\bar{A}}x = \bar{\bar{b}}$ avec $\bar{\bar{A}}$ inversible et donc $x = x_{lad} = \bar{\bar{A}}^{-1}\bar{\bar{b}}$ et que par conséquent $\bar{y} = \bar{y}_{lad} = \bar{A}\bar{\bar{A}}^{-1}\bar{\bar{b}} - \bar{b}$.

Cela signifie que si on utilise notre algorithme pour résoudre le problème des MD, on va arriver dans un dernier intervalle pour lequel dans (10), on a $X_1 = x_{lad}$, $V_1 = \bar{y}_{lad}$ et $W_1 = 0$.

4 Conclusions

Nous avons 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. Mais nous avons surtout proposé 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 oeuvre qui fournit la solution exacte (analytique) 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'obtenir a posteriori une solution robuste tout en évitant d'avoir à fixer a priori le seuil h dans le critère de Huber, l'objectif étant d'éliminer des observations aberrantes.

Références

[1] I. Barrodale and F.D.Roberts, "An improved algorithm for Discrete ℓ_1 linear Approximation," *SIAM J. Num. Analysis.*, 10, 5, 839-848, 1973.

[2] Y. Li et G.C. Arce, "A Maximum Likelihood Approach to Least Absolute Deviation Regression," *J. of Appl. Sign. Proc.*, 12, 1762-1769, 2004.

[3] P. Bloomfield et W.L. Steiger, *Least Absolute Deviations : Theory, Applications and Algorithms.* Birkhäuser, Boston, 1983.

[4] S.A. Ruzinsky and E.T. Olsen. L_1 and L_∞ Minimization via a Variant of Karmarkar's algorithm *IEEE-T-ASSP*, vol. 37, 2, pp. 245-253, Feb. 1989.

[5] P.J. Huber. Robust Statistics. *John Wiley and sons.*, New York, 1981.

[6] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization, IEEE Trans. Image Processing, 4(7) :932–946, July 1995.

[7] J.J. Fuchs, A new approach to robust linear regression. 14th IFAC World Congress, vol. H, pp. 427-432, july 99, Beijing.

[8] P.W. Holland and R.E. Welsh. Robust regression using iteratively reweighted least squares. *Comm. Stat.* A6, 813-828, 1977.

[9] S. Rosset et J. Zhu, "Piecewise Linear Regularized Solution Paths," *The Annals of Statistics.*, 35, 3, 1012-1030, 2007.

[10] R. Fletcher. Practical Methods of Optimization. *John Wiley and sons.*, 1987.