STAT 593
Robustness and Linear Models

Joseph Salmon

http://josephsalmon.eu

Télécom Paristech, Institut Mines-Télécom

&

University of Washington, Department of Statistics

(Visiting Assistant Professor)

Outline

Least Absolute Deviation

Equivariance

Least Trimmed Squares (LTS)

Table of Contents

Reminder on (Ordinary) Least squares, (O)LS

Model:
$\mathbf{y} \approx X\boldsymbol{\beta}^*$ where $\mathbf{y} \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \boldsymbol{\beta}^* \in \mathbb{R}^p$ (true coefficient)

A least square estimator is **any** solution of the following problem:

$$\hat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 := f(\boldsymbol{\beta})$$

$$\hat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^{n} \left[y_i - \langle x_i, \boldsymbol{\beta} \rangle \right]^2$$

Rem: Gaussian (-log)-likelihood leads to square formulation

Least Absolute Deviation (LAD)

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_1 := f(\boldsymbol{\beta})$$

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} |y_i - \langle x_i, \boldsymbol{\beta} \rangle|$$

Many properties, see Bloomfield and Steiger (1983) for instance for historical purspose

When $p = 1$, the estimator is

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} |y_i - x_i \boldsymbol{\beta}|$$

and one can find a solution with zero residuals, *i.e.*, $y_{i_0} = x_{i_0}\boldsymbol{\beta}$

Proof

First, one can simplify the problems to cases without any $x_i = 0$ by noticing that

$$\sum_{i=1}^{n} |y_i - x_i\boldsymbol{\beta}| \geq \sum_{i:x_i \neq 0} |y_i - x_i\boldsymbol{\beta}| + \sum_{i:x_i=0} |y_i|$$

Second, we assume "ab absurdum" that no solution achieves zero residuals. Ordering the slopes $\frac{y_1}{x_1} \leq \cdots \leq \frac{y_n}{x_n}$ one can assume that $\hat{\boldsymbol{\beta}}$, a LAD solution satisfies: $\hat{i} \in [n]$ s.t. $\hat{\boldsymbol{\beta}} \in \left(\frac{y_{\hat{i}}}{x_{\hat{i}}}, \frac{y_{\hat{i}+1}}{x_{\hat{i}+1}}\right)$

By Fermat's rule and hypothesis: $\displaystyle\sum_{i:\hat{\boldsymbol{\beta}} > \frac{y_i}{x_i}} |x_i| = \sum_{i:\hat{\boldsymbol{\beta}} < \frac{y_i}{x_i}} |x_i|$

One can check that $\tilde{\boldsymbol{\beta}} = \frac{y_{\hat{i}}}{x_{\hat{i}}}$, also satisfies the first order condition:

$$\sum_{i:\tilde{\boldsymbol{\beta}} > \frac{y_i}{x_i}} |x_i| - \sum_{i:\tilde{\boldsymbol{\beta}} < \frac{y_i}{x_i}} |x_i| + |x_{\hat{i}}| = \sum_{i:\hat{\boldsymbol{\beta}} > \frac{y_i}{x_i}} |x_i| - \sum_{i:\hat{\boldsymbol{\beta}} < \frac{y_i}{x_i}} |x_i| = 0$$

□

LAD in any dimension

Theorem

There exist at least one solution $\hat{\boldsymbol{\beta}}$ of the LAD for which $y_i = \langle\, x_i\,,\, \boldsymbol{\beta}\,\rangle$ for at least $\mathrm{rank}(X)$ indices.

LAD in any dimension

Theorem

There exist at least one solution $\hat{\beta}$ of the LAD for which $y_i = \langle x_i , \boldsymbol{\beta} \rangle$ for at least $\mathrm{rank}(X)$ indices.

Proof: this is provided in Th.1, Bloomfield and Steiger (1983). It works "ab absurdum": then there exist δ s.t. $\langle \delta , x_i \rangle = 0$ for indices with $y_i = \langle x_i , \boldsymbol{\beta} \rangle$ and $\langle \delta , x_i \rangle \neq 0$ for indices with $y_i \neq \langle x_i , \boldsymbol{\beta} \rangle$, then the objective is

$$\sum_{i : y_i \neq \langle x_i , \boldsymbol{\beta} \rangle} |y_i - \langle \boldsymbol{\beta} , x_i \rangle - t \langle \delta , x_i \rangle |$$

for the point $\boldsymbol{\beta} + t\delta$. With the previous lemma, one can create one more point that zeros the residual. This can be repeated except if one reaches $\mathrm{rank}(X)$ indices.

Table of Contents

Regression equivariance

Let T be an estimator of $\boldsymbol{\beta}^*$ (regression coeff.) based on $Z = (X, \mathbf{y})$

Definition

We say that T is **regression equivariant** when for any dataset (y, X) and any vector $v \in \mathbb{R}^p$, one has

$$T(X, y + Xv) = T(X, y) + v$$

Rem: a simple case is the OLS (full rank case)

$$(X^\top X)^{-1} X^\top (y + Xv) = (X^\top X)^{-1} X^\top y + v$$

Scale equivariance

Let T be an estimator of $\boldsymbol{\beta}^*$ (regression coeff.) based on $Z = (X, \mathbf{y})$

Definition

We say that T is **scale equivariant** when for any dataset (y, X) and any vector $c \in \mathbb{R}$, one has

$$T(X, c \cdot y) = c \cdot T(X, y)$$

Rem: a simple case is the OLS (full rank case)

$$(X^\top X)^{-1} X^\top (cy) = c(X^\top X)^{-1} X^\top y$$

Affine equivariance

Let T be an estimator of $\boldsymbol{\beta}^*$ (regression coeff.) based on $Z = (X, \mathbf{y})$

Definition

We say that T is **affine equivariant** when for any dataset (y, X) and any non-singular matrix $A \in \mathbb{R}^{p \times p}$, one has

$$T(XA, y) = A^{-1} T(X, y)$$

Rem: a simple case is the OLS (full rank case)

$$(A^\top X^\top X A)^{-1} (A)^\top X^\top (y) = A^{-1} (X^\top X)^{-1} X^\top y$$

Table of Contents

LTS Definition

Definition

For $h \in [n]$, the **Least Trimmed Squares** (LTS) estimator of order h is defined by

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{h} (r^2(\boldsymbol{\beta}))_{i:n},$$

where the vector $r^2(\boldsymbol{\beta}) = ((y_1 - \langle x_1 , \boldsymbol{\beta} \rangle)^2, \dots, (y_n - \langle x_n , \boldsymbol{\beta} \rangle)^2)$ represent the square **residuals** and $(r^2(\boldsymbol{\beta}))_{1:n} \leq \cdots \leq (r^2(\boldsymbol{\beta}))_{n:n}$ are the ordered statistics of the squared residuals

Rem: when $h < p$, LTS not uniquely defined

Rem: when $h = n$, LTS reduces to standard OLS

Alternative formulations

Set formulation: For $H \subset [n]$, we write
$Q(H, \boldsymbol{\beta}) = \|X_H \boldsymbol{\beta} - \mathbf{y}_H\|^2 = \sum_{i \in H} (y_i - \langle \boldsymbol{\beta}, x_i \rangle)^2$ then

$$(\hat{\boldsymbol{\beta}}, \hat{H}) \in \operatorname*{arg\,min}_{\substack{H \subset [n] : \# H = h \\ \boldsymbol{\beta} \in \mathbb{R}^p}} Q(H, \boldsymbol{\beta})$$

Alternative formulations

Set formulation: For $H \subset [n]$, we write
$Q(H, \boldsymbol{\beta}) = \|X_H \boldsymbol{\beta} - \mathbf{y}_H\|^2 = \sum_{i \in H} (y_i - \langle \boldsymbol{\beta}, x_i \rangle)^2$ then

$$(\hat{\boldsymbol{\beta}}, \hat{H}) \in \underset{\substack{H \subset [n] : \# H = h \\ \boldsymbol{\beta} \in \mathbb{R}^p}}{\arg\min} Q(H, \boldsymbol{\beta})$$

Binary variables formulation:

$$(\hat{\boldsymbol{\beta}}, \hat{w}) \in \underset{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ w \subset \mathbb{R}^n \\ \forall i \in [n], w_i \in \{0,1\} \\ \text{and } \sum_{i=1}^n w_i = h}}{\arg\min} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta}, x_i \rangle)^2$$

Rem: the later formulation is called a **Mixed Integer Programming** problem. Convex relaxation can be obtained by substituting $w_i \in [0, 1]$ to $w_i \in \{0, 1\}$, or optimization solver (like mosek, gurobi, etc.) can be considered.

Equivariance

Theorem

The LTS estimator is regression, scale and affine equivariant

Proof: consider the case where the data is $y + Xv$. Fix $H \in [n]$, as the optimal values in the LTS definition:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \sum_{i \in H} (y_i + \langle v , x_i \rangle - \langle \boldsymbol{\beta} , x_i \rangle)^2$$

Equivariance

Theorem

The LTS estimator is regression, scale and affine equivariant

Proof: consider the case where the data is $y + Xv$. Fix $H \in [n]$, as the optimal values in the LTS definition:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \sum_{i \in H} (y_i + \langle v , x_i \rangle - \langle \boldsymbol{\beta} , x_i \rangle)^2$$

$$\in \arg\min_{\boldsymbol{\beta}} \sum_{i \in H} (y_i - \langle \boldsymbol{\beta} - v , x_i \rangle)^2$$

Equivariance

Theorem

The LTS estimator is regression, scale and affine equivariant

Proof: consider the case where the data is $y + Xv$. Fix $H \in [n]$, as the optimal values in the LTS definition:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &\in \arg\min_{\boldsymbol{\beta}} \sum_{i \in H} (y_i + \langle\, v \,,\, x_i \,\rangle - \langle\, \boldsymbol{\beta} \,,\, x_i \,\rangle)^2 \\ &\in \arg\min_{\boldsymbol{\beta}} \sum_{i \in H} (y_i - \langle\, \boldsymbol{\beta} - v \,,\, x_i \,\rangle)^2 \\ &\in v + \arg\min_{\boldsymbol{\beta}} \sum_{i \in H} (y_i - \langle\, \boldsymbol{\beta} \,,\, x_i \,\rangle)^2\end{aligned}$$

Breakdown point: Donoho's definition

Definition: Breakdown point

For a dataset $Z = (X, \mathbf{y})$ where $X \in \mathbb{R}^{n \times p}$ corresponds to the design matrix and $\mathbf{y} \in \mathbb{R}^n$ to the observation vector, the **breakdown point** of a statistic T is:

$$\varepsilon^* = \varepsilon^*(T, Z) = \frac{m^*}{n + m^*}$$

where

$$m^* = \min \left\{ m : \sup_{\#Z' = m} \|T(Z \cup Z') - T(Z)\| = +\infty \right\}$$

Rem: ε-replacement variants often considered, see proof in Rousseeuw and Leroy (1987)

Breakdown point[1,2]

For simplicity we assume a classical full rank design assumption (so $p < n$).

Theorem

The breakdown point of any regression and permutation equivariant estimator is less than or equal to $\frac{n-p+1}{2n-p+1}$.

Rem: Asymptotically this is about a 50% breakdown point.

[1] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc., 1987, pp. xvi+329.

[2] D. L. Donoho. "Breakdown properties of multivariate location estimators". PhD thesis. Harvard University, 1982.

Proof

ab absurdum: assume $\exists B$ s.t.
$\sup_{\#Z'=n-p+1} \|T(Z' \cup Z) - T(Z)\| < B$
Up to a samples reordering, because $p-1$ vectors extracted among the row of X are included in a hyperplane, $\exists \mu \in \mathbb{R}^p$, with $\mu \neq 0$, s.t. $\langle \mu , x_1 \rangle = \cdots = \langle \mu , x_{p-1} \rangle = 0$.

Proof

ab absurdum: assume $\exists B$ s.t.
$\sup_{\#Z'=n-p+1} \|T(Z' \cup Z) - T(Z)\| < B$
Up to a samples reordering, because $p-1$ vectors extracted among the row of X are included in a hyperplane, $\exists \mu \in \mathbb{R}^p$, with $\mu \neq 0$, s.t. $\langle \mu , x_1 \rangle = \cdots = \langle \mu , x_{p-1} \rangle = 0$. Consider:

$$Z \cup Z' = \begin{pmatrix} x_1, & y_1 \\ \vdots & \vdots \\ x_{p-1}, & y_{p-1} \\ x_p, & y_p \\ \vdots & \vdots \\ x_n, & y_n \\ x_p, & y_p + \langle \mu , x_p \rangle \\ \vdots & \vdots \\ x_n, & y_n + \langle \mu , x_n \rangle \end{pmatrix}$$

Proof

ab absurdum: assume $\exists B$ s.t.
$\sup_{\#Z'=n-p+1} \|T(Z' \cup Z) - T(Z)\| < B$
Up to a samples reordering, because $p-1$ vectors extracted among the row of X are included in a hyperplane, $\exists \mu \in \mathbb{R}^p$, with $\mu \neq 0$, s.t. $\langle \mu , x_1 \rangle = \cdots = \langle \mu , x_{p-1} \rangle = 0$. Consider:

$$Z \cup Z' = \begin{pmatrix} x_1, & y_1 \\ \vdots & \vdots \\ x_{p-1}, & y_{p-1} \\ x_p, & y_p \\ \vdots & \vdots \\ x_n, & y_n \\ x_p, & y_p + \langle \mu , x_p \rangle \\ \vdots & \vdots \\ x_n, & y_n + \langle \mu , x_n \rangle \end{pmatrix} = \begin{pmatrix} x_1, & y_1 + \langle \mu , x_1 \rangle \\ \vdots & \vdots \\ x_{p-1}, & y_{p-1} + \langle \mu , x_{p-1} \rangle \\ x_p, & y_p \\ \vdots & \vdots \\ x_n, & y_n \\ x_p, & y_p + \langle \mu , x_p \rangle \\ \vdots & \vdots \\ x_n, & y_n + \langle \mu , x_n \rangle \end{pmatrix}$$

Proof (continued)

So by regression equivariance, reminding $Z = (X, \mathbf{y})$

$$T \begin{pmatrix} x_1, & y_1 + \langle\, \mu \,,\, x_1 \,\rangle \\ \vdots & \vdots \\ x_{p-1}, & y_{p-1} + \langle\, \mu \,,\, x_{p-1} \,\rangle \\ x_p, & y_p \\ \vdots & \vdots \\ x_n, & y_n \\ x_p, & y_p + \langle\, \mu \,,\, x_p \,\rangle \\ \vdots & \vdots \\ x_n, & y_n + \langle\, \mu \,,\, x_n \,\rangle \end{pmatrix} = T \begin{pmatrix} x_1, & y_1 \\ \vdots & \vdots \\ x_{p-1}, & y_{p-1} \\ x_p, & y_p - \langle\, \mu \,,\, x_p \,\rangle \\ \vdots & \vdots \\ x_n, & y_n - \langle\, \mu \,,\, x_n \,\rangle \\ x_p, & y_p \\ \vdots & \vdots \\ x_n, & y_n \end{pmatrix} + \mu$$

and then $T(Z \cup Z') = T(Z \cup Z'') + \mu$ for another dataset Z'' of size $n - p + 1$

Proof ending

By hypothesis: $\|T(Z \cup Z') - T(Z)\| \leq B$, but now one has also

$$\|T(Z \cup Z') - T(Z)\| = \|T(Z \cup Z'') + \mu - T(Z)\|$$

But Z'' being of size $n - p + 1$, then one has :

$$\|T(Z \cup Z'') - T(Z)\| \leq B$$

Since $\|\mu\|$ can be made arbitrarily large, leading to a contradiction.

Breakdown point[3,4]

Theorem

The (ε-contamination) breakdown point of the LTS is $\frac{h}{n+h}$. When $h = n - p + 1$, this reaches the largest bound for regression equivariant estimators, *i.e.*, $\frac{n-p+1}{2n-p+1}$

Rem: when n is large w.r.t. to p this is approximately 50%

[3] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc., 1987, pp. xvi+329.

[4] D. L. Donoho. "Breakdown properties of multivariate location estimators". PhD thesis. Harvard University, 1982.

Proof

Let $Z' = Z \cup \tilde{Z}$ the dataset, where one has added the h corrupted elements $(\tilde{x}_1, \tilde{y}_1), \ldots (\tilde{x}_h, \tilde{y}_h)$ (pick $h = n - p + 1$ to reach optimum)

Proof

Let $Z' = Z \cup \tilde{Z}$ the dataset, where one has added the h corrupted elements $(\tilde{x}_1, \tilde{y}_1), \ldots (\tilde{x}_h, \tilde{y}_h)$ (pick $h = n - p + 1$ to reach optimum)

To simplify the proof, we prove the lower bound for the Ridge version of the LTS estimator only:

$$(\hat{\boldsymbol{\beta}}, \hat{H}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, H : \# H = h}{\arg\min} \; Q(H, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|^2$$

$$\text{where} \quad Q(H, \boldsymbol{\beta}) = \left\| X'_H \boldsymbol{\beta} - \mathbf{y}'_H \right\|^2 = \sum_{i \in H} (y'_i - \langle \, \boldsymbol{\beta} \, , \, x'_i \, \rangle)^2$$

Proof continued

This means that for the Ridge version of the LTS, we prove that when one modifies h (or less) samples the estimator remains bounded.

Proof continued

This means that for the Ridge version of the LTS, we prove that when one modifies h (or less) samples the estimator remains bounded.

$$Q(H^*, 0) = \min_{H:\#H=h} Q(H, 0) = \sum_{i=1}^{h} y_{i:n}^2 \leq h \left\| y \right\|_\infty$$

Proof continued

This means that for the Ridge version of the LTS, we prove that when one modifies h (or less) samples the estimator remains bounded.

$$Q(H^*, 0) = \min_{H:\#H=h} Q(H, 0) = \sum_{i=1}^{h} y_{i:n}^2 \leq h \left\| y \right\|_\infty$$

Assume that $\left\| \boldsymbol{\beta} \right\|^2 \geq \frac{1+h\|y\|_\infty}{\lambda}$, then

$$\min_{H:\#H=h} Q(H, \boldsymbol{\beta}) + \lambda \left\| \boldsymbol{\beta} \right\|^2 \geq \lambda \left\| \boldsymbol{\beta} \right\|^2 \geq \lambda \frac{1 + h \left\| y \right\|_\infty}{\lambda} > Q(H^*, 0)$$

Proof continued

This means that for the Ridge version of the LTS, we prove that when one modifies h (or less) samples the estimator remains bounded.

$$Q(H^*, 0) = \min_{H:\#H=h} Q(H, 0) = \sum_{i=1}^{h} y_{i:n}^2 \leq h \left\| y \right\|_\infty$$

Assume that $\left\| \boldsymbol{\beta} \right\|^2 \geq \frac{1+h\|y\|_\infty}{\lambda}$, then

$$\min_{H:\#H=h} Q(H, \boldsymbol{\beta}) + \lambda \left\| \boldsymbol{\beta} \right\|^2 \geq \lambda \left\| \boldsymbol{\beta} \right\|^2 \geq \lambda \frac{1 + h \left\| y \right\|_\infty}{\lambda} > Q(H^*, 0)$$

Proof continued

This means that for the Ridge version of the LTS, we prove that when one modifies h (or less) samples the estimator remains bounded.

$$Q(H^*, 0) = \min_{H:\#H=h} Q(H, 0) = \sum_{i=1}^{h} y_{i:n}^2 \leq h \|y\|_\infty$$

Assume that $\|\boldsymbol{\beta}\|^2 \geq \frac{1+h\|y\|_\infty}{\lambda}$, then

$$\min_{H:\#H=h} Q(H, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|^2 \geq \lambda \|\boldsymbol{\beta}\|^2 \geq \lambda \frac{1 + h \|y\|_\infty}{\lambda} > Q(H^*, 0)$$

Now since $\min_{\boldsymbol{\beta}, H:\#H=h} Q(H, \boldsymbol{\beta}) \leq Q(H^*, 0)$, one needs to have $\left\|\hat{\boldsymbol{\beta}}\right\|^2 \leq \frac{1+h\|y\|_\infty}{\lambda}$, a bound that does not depend on the $\tilde{x}_i, \tilde{y}_i$ □

Optimization for LTS : Mixed Integer Programming

Generic approach; requires fast solvers like gurobi, mosek, cplex, etc.
Ingredients:

- Convex relaxation : convexify the binary constraints

$$P \quad = \quad \min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p \\ w\subset\mathbb{R}^n \\ \forall i\in[n], w_i\in\{0,1\} \\ \text{and } \sum_{i=1}^n w_i = h}} \quad \sum_{i=1}^{n} w_i(y_i - \langle\, \boldsymbol{\beta} \,,\, x_i \,\rangle)^2$$

- If a solution $\hat{w}$ of P has integer values stop: the global optimal solution has been found

Rem: $P^{\mathrm{cvx}} \leq P$ (lower bound on the optimal value)

Optimization for LTS : Mixed Integer Programming

Generic approach; requires fast solvers like gurobi, mosek, cplex, etc.

Ingredients:

- Convex relaxation : convexify the binary constraints

$$P^{\mathrm{cvx}} = \min_{\substack{\boldsymbol{\beta}\in\mathbb{R}^p \\ w\subset\mathbb{R}^n \\ \forall i\in[n], w_i\in[0,1] \\ \text{and } \sum_{i=1}^n w_i = h}} \sum_{i=1}^n w_i(y_i - \langle \boldsymbol{\beta} , x_i \rangle)^2$$

- If a solution $\hat{w}$ of P has integer values stop: the global optimal solution has been found

Rem: $P^{\mathrm{cvx}} \leq P$ (lower bound on the optimal value)

Branch and bound

Otherwise: "branch and bound", $\exists i_0 \in [n]$ such that $w_{i_0} \in]0,1[$ so solve two MIP problems:

$$P_l = \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ w \subset \mathbb{R}^n \\ \forall i \in [n], w_i \in \{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0} = 0}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta} , x_i \rangle)^2$$

$$P_r = \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ w \subset \mathbb{R}^n \\ \forall i \in [n], w_i \in \{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0} = 1}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta} , x_i \rangle)^2$$

The variable i_0 is called a **branching** variable

Branch and bound

Otherwise: "branch and bound", $\exists i_0 \in [n]$ such that $w_{i_0} \in]0,1[$ so solve two MIP problems:

$$P_l = \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ w \subset \mathbb{R}^n \\ \forall i \in [n], w_i \in \{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0} = 0}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta}, x_i \rangle)^2$$

$$P_r = \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ w \subset \mathbb{R}^n \\ \forall i \in [n], w_i \in \{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0} = 1}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta}, x_i \rangle)^2$$

The variable i_0 is called a **branching** variable

Now one has $P = \min(P_l, P_r)$, and one can solve recursively the problems P_r and P_l by proceeding similarly (use a **search tree**, and in general no need to solve the 2^n sub-problems)

Branch and bound

Otherwise: "branch and bound", $\exists i_0 \in [n]$ such that $w_{i_0} \in]0,1[$ so solve two MIP problems:

$$P_l = \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ w \subset \mathbb{R}^n \\ \forall i \in [n], w_i \in \{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0} = 0}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta} , x_i \rangle)^2$$

$$P_r = \min_{\substack{\boldsymbol{\beta} \in \mathbb{R}^p \\ w \subset \mathbb{R}^n \\ \forall i \in [n], w_i \in \{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0} = 1}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta} , x_i \rangle)^2$$

The variable i_0 is called a **branching** variable

Now one has $P = \min(P_l, P_r)$, and one can solve recursively the problems P_r and P_l by proceeding similarly (use a **search tree**, and in general no need to solve the 2^n sub-problems)

Rem: other useful bounds are $P^{\text{cvx}} \leq \min(P_l^{\text{cvx}}, P_r^{\text{cvx}}) \leq P$

Branch and bound

Otherwise: "branch and bound", $\exists i_0 \in [n]$ such that $w_{i_0} \in]0,1[$ so solve two MIP problems:

$$P_l = \min_{\substack{\beta\in\mathbb{R}^p \\ w\subset\mathbb{R}^n \\ \forall i\in[n], w_i\in\{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0}=0}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta}, x_i \rangle)^2 \qquad P_r = \min_{\substack{\beta\in\mathbb{R}^p \\ w\subset\mathbb{R}^n \\ \forall i\in[n], w_i\in\{0,1\} \\ \sum_{i=1}^n w_i = h \\ w_{i_0}=1}} \sum_{i=1}^n w_i (y_i - \langle \boldsymbol{\beta}, x_i \rangle)^2$$

The variable i_0 is called a **branching** variable

Now one has $P = \min(P_l, P_r)$, and one can solve recursively the problems P_r and P_l by proceeding similarly (use a **search tree**, and in general no need to solve the 2^n sub-problems)

<u>Rem</u>: other useful bounds are $P^{\text{cvx}} \leq \min(P_l^{\text{cvx}}, P_r^{\text{cvx}}) \leq P$

<u>Rem</u>: upper bounds can be obtained by finding feasible points (*e.g.*, rounding)

Fast LTS

Simple alternative: iterative procedure Rousseeuw and Van Driessen(2006)

Algorithm: Fast LTS

---

input : h, max. iterations $t_{\max}$, stopping criterion ε
init : $H^0, \boldsymbol{\beta}^0$
for $1 \leq t \leq t_{\max}$ **do**

  Break if stopping criterion smaller than ε

$$H^{t+1} \leftarrow \underset{H:\#H=h}{\arg\min} \left\| X_H \boldsymbol{\beta}^t - \mathbf{y}_H \right\|^2$$

$$\boldsymbol{\beta}^{t+1} \leftarrow \underset{\boldsymbol{\beta}}{\arg\min} \left\| X_{H^{t+1}} \boldsymbol{\beta} - \mathbf{y}_{H^{t+1}} \right\|^2$$

return $\boldsymbol{\beta}^t, H^t$

Fast LTS

Simple alternative: iterative procedure Rousseeuw and Van Driessen(2006)

Algorithm: FAST LTS

input : h, max. iterations $t_{\max}$, stopping criterion ε
init : $H^0, \boldsymbol{\beta}^0$
for $1 \leq t \leq t_{\max}$ **do**
  Break if stopping criterion smaller than ε

$$H^{t+1} \leftarrow \underset{H:\#H=h}{\arg\min} \left\| X_H \boldsymbol{\beta}^t - \mathbf{y}_H \right\|^2$$

$$\boldsymbol{\beta}^{t+1} \leftarrow \underset{\boldsymbol{\beta}}{\arg\min} \left\| X_{H^{t+1}} \boldsymbol{\beta} - \mathbf{y}_{H^{t+1}} \right\|^2$$

return $\boldsymbol{\beta}^t, H^t$

Rem: $Q(H^{t+1}, \boldsymbol{\beta}^{t+1}) \leq Q(H^{t+1}, \boldsymbol{\beta}^t) \leq Q(H^t, \boldsymbol{\beta}^t)$

Another simpler alternative : Fast LTS

- the update

$$H^{t+1} \leftarrow \underset{H:\#H=h}{\arg\min} \left\| X_H \boldsymbol{\beta}^t - \mathbf{y}_H \right\|^2$$

can be obtained in a closed form by sorting; cost$=$ $O(n \log(n))$ or less if h is small (use: `np.partition` in `numpy`)

Another simpler alternative : Fast LTS

- the update

$$H^{t+1} \leftarrow \underset{H:\#H=h}{\arg\min} \left\| X_H \boldsymbol{\beta}^t - \mathbf{y}_H \right\|^2$$

  can be obtained in a closed form by sorting; cost= $O(n \log(n))$ or less if h is small (use: `np.partition` in `numpy`)

- inner solver needed for the second update:

$$\boldsymbol{\beta}^{t+1} \leftarrow \underset{\boldsymbol{\beta}}{\arg\min} \left\| X_{H^{t+1}} \boldsymbol{\beta} - \mathbf{y}_{H^{t+1}} \right\|^2$$

  A second stopping criteria is then needed; possibly do not solve too precisely the problem at each step

Another simpler alternative : Fast LTS

- the update

$$H^{t+1} \leftarrow \underset{H:\#H=h}{\arg\min} \left\| X_H \boldsymbol{\beta}^t - \mathbf{y}_H \right\|^2$$

  can be obtained in a closed form by sorting; cost= $O(n \log(n))$ or less if h is small (use: `np.partition` in `numpy`)

- inner solver needed for the second update:

$$\boldsymbol{\beta}^{t+1} \leftarrow \underset{\boldsymbol{\beta}}{\arg\min} \left\| X_{H^{t+1}} \boldsymbol{\beta} - \mathbf{y}_{H^{t+1}} \right\|^2$$

  A second stopping criteria is then needed; possibly do not solve too precisely the problem at each step
- initialization is tricky (*e.g.*, similar to K-means issues), might use several initialization

Summary on optimizing LTS

2 directions:

- Mixed Integer Programming
  - pros: bounds / certificate for optimality
  - cons: more complex to implement, need of specific solvers
- Alternate minimization
  - pros: simple to implement
  - cons: initialization, no guarantee (only convergence to local minimum)

Rem: hybrid method could be useful, as MIP can benefit from a nicer initialization (through nicer upper bounds)
Rem: "continuation" method can also be proposed, *i.e.*, start by small h (fast to solve) and then increase h progressively

LTS extensions through regularization[5]

Adapt to high dimensional constraints using regularization:

$$(\hat{\boldsymbol{\beta}}, \hat{H}) \in \underset{\substack{H \subset [n]: \# H = h \\ \boldsymbol{\beta} \in \mathbb{R}^p}}{\arg\min} \; Q(H, \boldsymbol{\beta}) + h\lambda \, \mathrm{pen}(\boldsymbol{\beta})$$

where $Q(H, \boldsymbol{\beta}) = \|X_H \boldsymbol{\beta} - \mathbf{y}_H\|^2 = \sum_{i \in H} (y_i - \langle \, \boldsymbol{\beta} \, , \, x_i \, \rangle)^2$

- Ridge penalty (as seen earlier): $\mathrm{pen}(\boldsymbol{\beta}) = \|\beta\|^2$
- Lasso penalty for sparsity enforcing: $\mathrm{pen}(\boldsymbol{\beta}) = \|\beta\|_1$
- etc.

<u>Rem</u>: such approaches loose regression equivariance by enforcing specific constraints on the targeted solution (*e.g.*, sparsity)

[5] A. Alfons, C. Croux, and S. Gelper. "Sparse least trimmed squares regression for analyzing high-dimensional large data sets". In: *Ann. Appl. Stat.* 7.1 (2013), pp. 226–248.

References and supplementary material

- For extensions to joint estimation of β and noise level σ *cf.* Ch. 6, Maronna *et al.* (2006)

  Example : consider for $\hat{\beta}$ being the LTS

  $$\hat{\sigma} = \frac{1}{h}\sum_{i=1}^{h}(r^2(\hat{\beta}))_{i:n},$$

- Heteroscedastic models (case where the noise level differs for each observations) Ch. 6, Maronna *et al.* (2006)

- branch and bound:
  https://web.stanford.edu/class/ee392o/bb.pdf

References I

- Alfons, A., C. Croux, and S. Gelper. "Sparse least trimmed squares regression for analyzing high-dimensional large data sets". In: *Ann. Appl. Stat.* 7.1 (2013), pp. 226–248.
- Bloomfield, P. and W. L. Steiger. *Least absolute deviations*. Vol. 6. Progress in Probability and Statistics. Theory, applications, and algorithms. Birkhäuser Boston, Inc., Boston, MA, 1983, pp. xiv+349.
- Donoho, D. L. "Breakdown properties of multivariate location estimators". PhD thesis. Harvard University, 1982.
- Maronna, R. A., R. D. Martin, and V. J. Yohai. *Robust statistics: Theory and methods*. Chichester: John Wiley & Sons, 2006.
- Rousseeuw, P. J. and A. M. Leroy. *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc., 1987, pp. xvi+329.

References II

- Rousseeuw, P. J. and K. Van Driessen. “Computing LTS Regression for Large Data Sets”. In: *Data Mining and Knowledge Discovery* 12.1 (2006), pp. 29–45.