Accounting for outliers in optimal subsampling methods

Laura Deldossi[1] Elena Pesce[2] Chiara Tommasi[3]

[1]Department of Statistical Sciences,
Università Cattolica del Sacro Cuore, Milan, Italy

[2]Swiss Re Institute, Zurich, Switzerland

[3]Department of Economics, Management and Quantitative Methods
University of Milan, Italy

mODa13 Workshop
Southampton (UK), 9-14 July 2023

1. Motivation of the work

Nowadays, advances in technology have brought the ability to collect, transfer and store large datasets.

Data reduction may help in reducing not only the computational burden but also the costs of querying the Big Dataset.

Among various subsampling techniques, the **design inspired subsampling methods** attracted great interest in the last few years. A review of these methods is available in Yu et al. (2023), who classify them according to the different kinds of design adopted:

- optimal design
- orthogonal design
- space filling design

Sample selection based on the theory of optimal design

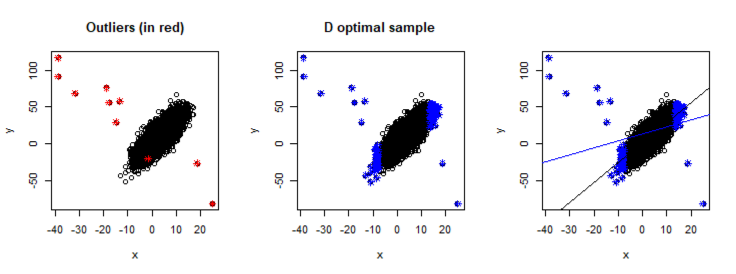
The theory of optimal design is a guide to draw a subsample containing the most informative observations to provide accurate statistical inference with minimum cost.
Optimal subsamples lie on the boundary of the region.

Drawback: Big Datasets usually are the result of passive observation thus abnormal values may be present.

The theory of optimal design is a guide to draw a subsample containing the most informative observations to provide accurate statistical inference with minimum cost.
Optimal subsamples lie on the boundary of the region.

Drawback: Big Datasets usually are the result of passive observation thus abnormal values may be present.

Example



OUR GOAL

To select a subsample to produce an efficient parameter estimate (or an accurate prediction) for the model generating the whole dataset apart from a few outliers.

2. Framework and notation

Assume that N independent responses have been generated from a super-population model

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, N,$$

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$
- $\mathbf{x}_i^\top = (1, \tilde{\mathbf{x}}_i^\top)$ with $\tilde{\mathbf{x}}_i = (x_{i1}, \ldots, x_{ik})^\top$
- ε_i iid random errors with zero mean and equal variance σ^2.

- We assume that N (the number of items of the Big Dataset) is much larger than k (the number of features), for this reason we do not consider data reduction in the features domain (dimensionality reduction techniques).

- $U = \{1, \ldots, N\}$ denotes the population of units under study

- $s_n = \{i_1, \ldots, i_n\} \subseteq U$ denotes a **subsample** without replications of size n from U

- Given the sample $s_n = \{i_1, \dots, i_n\}$, the **least squares estimator** of $\boldsymbol{\beta}$ is

$$\hat{\beta} = \hat{\beta}(s_n) = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} = \left(\sum_{\ell=1}^{N} \mathbf{x}_\ell \mathbf{x}_\ell^T I_\ell\right)^{-1} \sum_{\ell=1}^{N} \mathbf{x}_\ell \, Y_\ell \, I_\ell$$

where:
$\mathbf{X}$ is the $n \times (k+1)$ matrix whose rows are $\mathbf{x}_i^\top$ for $i \in s_n$
$\mathbf{Y} = (Y_{i_1}, \dots, Y_{i_n})^\top$ is the vector of responses of the units in s_n
and

$$I_\ell = \begin{cases} 1 & \text{if } \ell \in s_n \\ 0 & \text{otherwise} \end{cases}, \qquad \text{with } \ell = 1, \dots, N$$

is the **sample inclusion indicator**.

Precise estimation of the parameters

Precision measure

When the inferential goal is to get a **precise estimate** of β, a sample s_n should be selected drawing the n observations with the smallest generalized variance of $\hat{\beta}$: $\sigma^2|\mathbf{X}^\top\mathbf{X}|^{-1}$ or with the **largest** determinant of the **precision matrix**: $\sigma^{-2}|\mathbf{X}^\top\mathbf{X}|$.

D-optimal subsampling

$$s_n^D = \underset{s_n=\{I_1,\ldots,I_N\}}{\arg\sup} \left| \sum_{\ell=1}^{N} \mathbf{x}_\ell \mathbf{x}_\ell^\top I_\ell \right|, \quad I_\ell = \begin{cases} 1 & \text{if } \ell \in s_n \\ 0 & \text{otherwise} \end{cases}$$

A commonly applied algorithm to determine the D-optimal sample is the well known **exchange algorithm** (Chp. 12 in Atkinson et al. (2007)).

Algorithm 1: Exchange algorithm for D-optimality

Algorithm 1 Exchange Algorithm for D-optimality

Require: Design matrix $\boldsymbol{X}$, sample size n, initial sample $s_n^{(0)}$, t_{max}, $\tilde{N}$
Ensure: D-optimal sample

1: Set $t = 0$
2: **while** $t < t_{max}$ **do**
3: Select randomly $\tilde{N}$ units from $\left\{U - s_n^{(t)}\right\}$ to form the set of candidate points for the exchange, $\mathcal{C}^{(t)}$
4: Select from $\mathcal{C}^{(t)}$ the observation $j_a = \arg\max_{j \in \mathcal{C}^{(t)}} \boldsymbol{x}_j^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j$
5: Add unit j_a to $s_n^{(t)}$ to form the augmented sample $s_{n+1}^{(t)}$ of size $n+1$
6: From $s_{n+1}^{(t)}$ identify the unit with the smallest prediction variance $i_m = \arg\min_{i \in s_{n+1}^{(t)}} h_{ii}$
7: Remove unit i_m from $s_{n+1}^{(t)}$ to obtain the updated sample $s_n^{(t+1)}$
8: Set $t = t + 1$
9: **end while**

Augmentation step (step 5)

add the point $\mathbf{x}_{j_a}$ that provides the **maximum increase** in the determinant of the precision matrix. This is the point with the **largest leverage score**.

Deletion step (step 7)

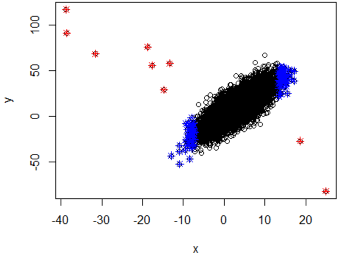
delete the point $\mathbf{x}_{i_m}$ that provides the **minimum decrease** in the determinant of the precision matrix, that is the unit with the **smallest leverage score**.

High leverage points

The units with the largest leverage score can be **good** or **bad**:

- they are good (blue points) when the related response is not an outlier and thus their inclusion would reduce the variance of the parameters' estimates;

- they are bad (red points) when they are associated to an "abnormal" response and thus it might alter the model fitted by the bulk of the data.



According to Hoaglin and Welsch (1978) an observation $\mathbf{x}_i$ with $i = 1, \ldots, n$ is called an *high leverage point* when

$$h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i > \nu_1 (k+1)/n$$

where ν_1 is a tuning parameter usually set equal to 2.

3. D-optimal sample without high leverage points/outliers

To construct a *D*-optimal sample avoiding (as much as possible) high leverage points/outliers, we propose two algorithms

- Algorithm 2, which is a **non-informative** method (not based on the response observations) and produces D-optimal samples without high leverage points;
- Algorithm 3, which is an **informative** method (based on the response observations) and gives D-optimal samples without outliers.

Modifications of the exchange algorithm

- switching the augmentation and deletion steps;
- changing the set $\mathcal{C}^{(t)}$ where the observation to be added is searched.

Algorithm 2: Non-informative D-optimal sample without high leverage points

Algorithm 2 Non-informative D-optimal sample without high leverage points

Require: Design matrix $\boldsymbol{X}$, sample size n, initial sample $s_n^{(0)}$, ν_1, t_{max}, $\tilde{N}$
Ensure: D-optimal sample without high leverage points

1: Set $t = 0$
2: **while** $t < t_{max}$ **do**
3: Identify the unit $i_m = \arg\min_{i \in s_n^{(t)}} h_{ii}$
4: From (3), compute the inverse of the information matrix without i_m: $(\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^{-})^{-1}$
5: Select randomly $\tilde{N}$ units from $\left\{U - s_n^{(t)}\right\}$
6: From (4), compute $h_{i_m i_m}(\boldsymbol{x}_j)$ $(j = 1, \ldots, \tilde{N})$, to identify the set of candidate points $\mathcal{C}^{(t)} = \left\{j : \; h_{i_m i_m} < h_{i_m i_m}(\boldsymbol{x}_j) < \nu_1 \frac{k+1}{n}\right\}$
7: Select from $\mathcal{C}^{(t)}$ the observation $j_a = \arg\max_{j \in \mathcal{C}^{(t)}} \boldsymbol{x}_j^{\top} (\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^{-})^{-1} \boldsymbol{x}_j$
8: Update $s_n^{(t)}$ by replacing unit i_m with j_a, to form $s_n^{(t+1)}$
9: Set $t = t + 1$
10: **end while**

Deletion steps (step 3-4)

delete the unit $\mathbf{i_m}$ with the **smallest leverage score**, thus obtaining a reduced sample of size $n - 1$.

Augmentation steps (step 7-8)

add the unit j_a in $\mathcal{C}^{(t)}$ that provides the **largest leverage score** when exchanged with the unit i_m .

We provide also an algorithm to find out an initial sample $s_n^{(0)}$ in the bulk of the data.

Analytical details regarding step 6 of Algorithm 2

Theorem 1

Theorem 1 *Let* ${}_j\boldsymbol{X}_t$ *be the design matrix obtained from* $\boldsymbol{X}_t$ *exchanging* $\boldsymbol{x}_{i_m}$ *with* $\boldsymbol{x}_j$*, then*

$$h_{i_m i_m}(\boldsymbol{x}_j) = \boldsymbol{x}_j^\top \left({}_j\mathbf{X}_t^\top {}_j\mathbf{X}_t\right)^{-1} \boldsymbol{x}_j \qquad (4)$$

where

$$\left({}_j\mathbf{X}_t^\top {}_j\mathbf{X}_t\right)^{-1} = (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} - (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \frac{\boldsymbol{A}}{d} (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1}, \qquad (5)$$

with

$$\begin{aligned}\boldsymbol{A} = {} & \boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j \,(\boldsymbol{x}_j \boldsymbol{x}_{i_m}^\top + \boldsymbol{x}_{i_m} \boldsymbol{x}_j^\top) + [1 - \boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_{i_m}]\, \boldsymbol{x}_j \boldsymbol{x}_j^\top \\ & - [1 + \boldsymbol{x}_j^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j]\, \boldsymbol{x}_{i_m} \boldsymbol{x}_{i_m}^\top;\end{aligned}$$

$$d = [1 - \boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_{i_m}]\,[1 + \boldsymbol{x}_j^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j] + [\boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j]^2.$$

Algorithm 3: Informative D-optimal sample without outliers

Algorithm 3 Informative optimal subsample without outliers

Require: Dataset $\boldsymbol{D}$, sample size n, initial sample $s_n^{(0)}$, ν_1, t_{max}, $\tilde{N}$
Ensure: Informative D-optimal sample without outliers

1: Set $t = 0$
2: **while** $t < t_{max}$ **do**
3: Identify the unit $i_m = \arg\min_{i \in s_n^{(t)}} h_{ii}$
4: From (3), compute the inverse of the information matrix without i_m: $(\boldsymbol{X}_t^{-\top}\boldsymbol{X}_t^{-})^{-1}$
5: Select randomly $\tilde{N}$ units from $\left\{U - s_n^{(t)}\right\}$
6: From (4), compute $h_{i_m i_m}(\boldsymbol{x}_j)$ $(j = 1, \ldots, \tilde{N})$, to identify the set of candidate points $\mathcal{C}^{(t)}$ according to (2)
7: Select from $\mathcal{C}^{(t)}$ the observation $j_a = \arg\max_{j \in \mathcal{C}^{(t)}} \boldsymbol{x}_j^\top(\boldsymbol{X}_t^{-\top}\boldsymbol{X}_t^{-})^{-1}\boldsymbol{x}_j$
8: Compute Cook's distance for unit j_a:

$$C_{j_a} = \frac{\left(Y_{j_a} - \hat{Y}_{j_a}\right)^2}{(k+1)\,\hat{\sigma}^2} \cdot \frac{h_{i_m i_m}(\boldsymbol{x}_{j_a})}{(1 - h_{i_m i_m}(\boldsymbol{x}_{j_a}))^2}$$

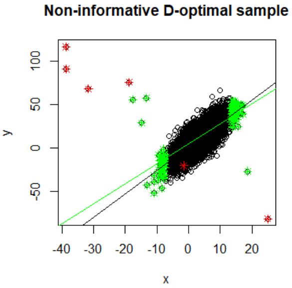
9: **if** $C_{j_a} < 4/n$ **then**
10: Update $s_n^{(t)}$ by replacing unit i_m with j_a, to form $s_n^{(t+1)}$
11: Set $t = t + 1$
12: **else**
13: reject the exchange and go back to step 5
14: **end if**
15: **end while**

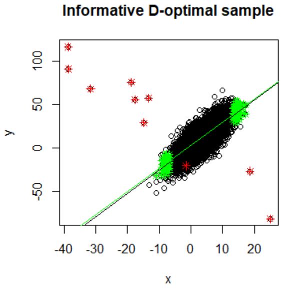
Cook's distance

$$C_i = \frac{\left(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)}\right)^\top\left(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)}\right)}{(k+1)\,\hat{\sigma}^2}$$

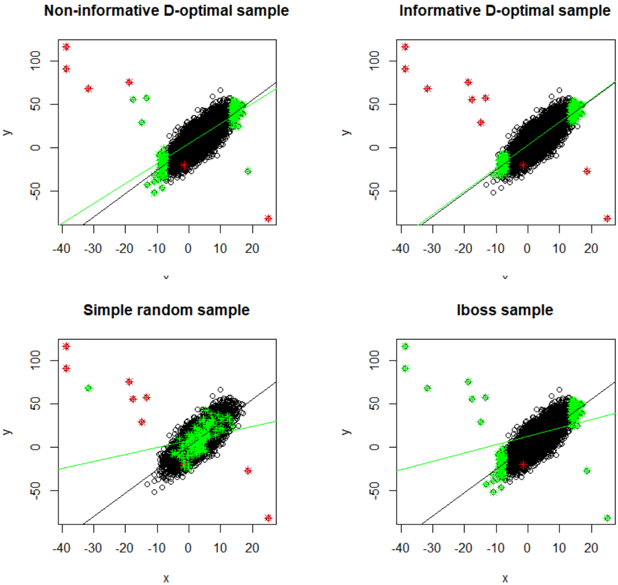
where $\hat{\boldsymbol{Y}}_{(i)}$ is the fit without the i-th unit. It measures how much all of the fitted values in the model **change** when the i-th data point is deleted. See Chatterjee et al (1986).

Example (follows)





Example (follows)



4. Optimal subsampling to get accurate prediction

Prediction accuracy

When the inferential goal is to get **accurate predictions** on a set of values $\mathcal{X}_0 = \{\mathbf{x}_{01}, \ldots, \mathbf{x}_{0N_0}\}$, we should select the observations minimizing the **overall prediction variance**:

$$\begin{aligned}\sum_{i=1}^{N_0} MSPE(\hat{Y}_{0i}|\mathbf{x}_{0j}, \mathbf{X}) &= \sum_{i=1}^{N_0} \mathrm{E}[(\hat{Y}_{0i} - \mu_{0i})^2|\mathbf{x}_{0j}, \mathbf{X}] = \\ \sigma^2 \cdot \mathrm{Trace}[\mathbf{X}_0(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}_0^\top] &= \sigma^2 \cdot \mathrm{Trace}[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}_0^\top \mathbf{X}_0],\end{aligned}$$

where $\hat{Y}_{0j} = \mathbf{x}_{0i}^T \hat{\beta}$ is the prediction of $\mu_{0i} = E(Y_{0i}|\mathbf{x}_{0i})$ and $\boldsymbol{X}_0$ is the $N_0 \times k$ matrix whose i-th row is $\boldsymbol{x}_{0i}^\top$, $i = 1, \ldots, N_0$.

I-optimal subsampling

$$s_n^I = \underset{s_n = \{I_1, \ldots, I_N\}}{\arg\inf} \ \mathrm{Trace}\left[\left(\sum_{\ell=1}^{N} \mathbf{x}_\ell \mathbf{x}_\ell^\top I_\ell\right)^{-1} \mathbf{X}_0^\top \mathbf{X}_0\right], \quad I_\ell = \begin{cases} 1 & \text{if } \ell \in s_n \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 4: Exchange algorithm for the I-optimality

Algorithm 4 Non-informative I-optimal sample without high leverage points

Require: Design matrix $\boldsymbol{X}$, sample size n, initial sample $s_n^{(0)}$, prediction-set $\mathcal{X}_0 = \{\boldsymbol{x}_{01}, \ldots, \boldsymbol{x}_{0N_0}\}$, ν_1, t_{max}, $\tilde{N}$

Ensure: I-optimal sample without high leverage points

1: Set $t = 0$

2: **while** $t < t_{max}$ **do**

3: Identify the unit

$$i_m = \arg\min_{i \in s_n^{(t)}} \frac{\boldsymbol{x}_i^\top \left(\boldsymbol{X}_t^\top \boldsymbol{X}_t\right)^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0 \left(\boldsymbol{X}_t^\top \boldsymbol{X}_t\right)^{-1} \boldsymbol{x}_i}{1 - \boldsymbol{x}_i^T (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_i}$$

4: From (3), compute the inverse of the information matrix without i_m: $(\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^{-})^{-1}$

5: Select randomly $\tilde{N}$ units from $\left\{U - s_n^{(t)}\right\}$

6: From (4) and (9), compute $h_{i_m i_m}(\boldsymbol{x}_j)$ and $\tilde{h}_{i_m i_m}(\boldsymbol{x}_j)$ $(j = 1, \ldots, \tilde{N})$, to identify the set of candidate points $\mathcal{C}^{(t)}$ according to (8)

7: Select from $\mathcal{C}^{(t)}$ the observation

$$j_a = \arg\max_{j \in \mathcal{C}^{(t)}} \frac{\boldsymbol{x}_j^\top \left(\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^{-}\right)^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0 \left(\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^{-}\right)^{-1} \boldsymbol{x}_j}{1 + \boldsymbol{x}_j^T (\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^{-})^{-1} \boldsymbol{x}_j}$$

8: Update $s_n^{(t)}$ by replacing unit i_m with j_a, to form $s_n^{(t+1)}$

9: Set $t = t + 1$

10: **end while**

Deletion steps (step 3-4)

delete the unit i_m whose omission minimises the increment in the overall mean squared prediction error.

Augmentation steps (step 7-8)

add the unit j_a in $\mathcal{C}^{(t)}$ that maximises the decrease in the overall mean squared prediction error

Candidate points set

$$\mathcal{C}^{(t)} = \left\{ j: \ \tilde{h}_{i_m i_m}(\boldsymbol{x}_j) > \tilde{h}_{i_m i_m} \cap h_{i_m i_m}(\boldsymbol{x}_j) < \nu_1 \frac{k+1}{n} \right\}, \tag{8}$$

where

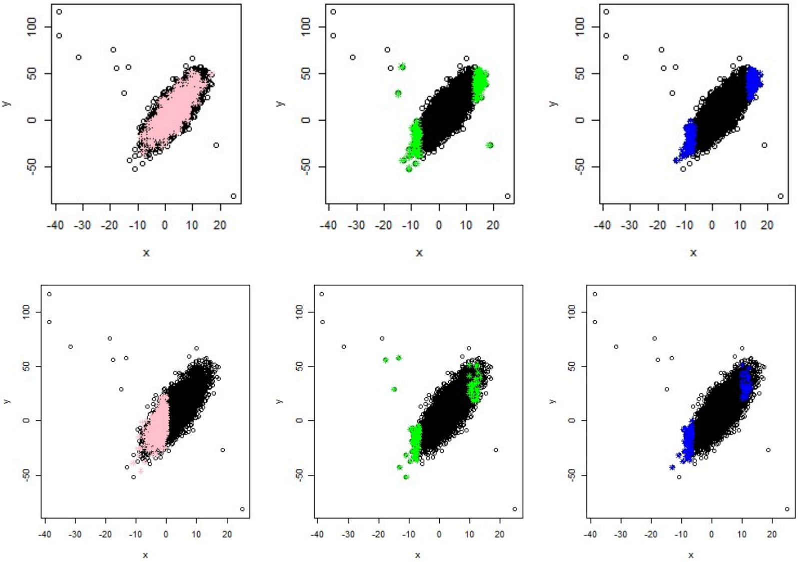
$$\tilde{h}_{ii} = \frac{\boldsymbol{x}_i^\top \, (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \, \boldsymbol{X}_0^\top \boldsymbol{X}_0 \, (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \, \boldsymbol{x}_i}{1 - \boldsymbol{x}_i^T (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_i},$$

$$\tilde{h}_{i_m i_m}(\boldsymbol{x}_j) = \frac{\boldsymbol{x}_j^T \, ({}_j\mathbf{X}_t^\top {}_j\mathbf{X}_t)^{-1} \, \boldsymbol{X}_0^\top \boldsymbol{X}_0 \, ({}_j\mathbf{X}_t^\top {}_j\mathbf{X}_t)^{-1} \, \boldsymbol{x}_j}{1 - \boldsymbol{x}_j^T ({}_j\mathbf{X}_t^\top {}_j\mathbf{X}_t)^{-1} \boldsymbol{x}_j}, \tag{9}$$

${}_j\boldsymbol{X}_t$ is the matrix obtained from $\boldsymbol{X}_t$ by exchanging $\boldsymbol{x}_{i_m}$ with $\boldsymbol{x}_j$ and $({}_j\mathbf{X}_t^\top {}_j\mathbf{X}_t)^{-1}$ can be computed from Equation (5).

Example (follows)

Prediction set $\mathcal{X}_0$ **Non-informative I-opt.** **Informative I-opt.**



5. A simulation study

$H \times S$ datasets ($H = 30$ and $S = 50$) of size $N = 10^6$, each one including $N_{out} = 500$ high leverage points/outliers are simulated:

$$ {}_h\boldsymbol{D}_s = [{}_h\boldsymbol{X}, {}_h\boldsymbol{y}_s], \quad h = 1, \dots, H, \; s = 1, \dots, S, \quad \text{where} $$

$$ {}_h\boldsymbol{X} = \begin{bmatrix} 1 & {}_h\tilde{\boldsymbol{x}}_1^\top \\ \vdots & \vdots \\ 1 & {}_h\tilde{\boldsymbol{x}}_N^\top \end{bmatrix} = \begin{bmatrix} 1 & {}_hx_{11} & \dots & {}_hx_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & {}_hx_{N1} & \dots & {}_hx_{Nk} \end{bmatrix} $$

$$ \swarrow \quad \downarrow \quad \searrow $$

$$ {}_h\boldsymbol{Y}_1 = \begin{bmatrix} {}_hY_{1,1} \\ \vdots \\ {}_hY_{1,N} \end{bmatrix}, \cdots, {}_h\boldsymbol{Y}_S = \begin{bmatrix} {}_hY_{S,1} \\ \vdots \\ {}_hY_{S,N} \end{bmatrix} $$

Simulated design matrix

Specifically, $k = 10$ and ${}_h\tilde{\mathbf{x}}_i = (x_{i1}, \ldots, x_{i10})^\top$, for $i = 1, \ldots, N$, are generated as follows:

- x_{i1}, x_{i2} and x_{i3} are independently distributed as $U(0, 5)$;
- $(x_{i4}, x_{i5}, x_{i6}, x_{i7})^\top$ is distributed as a multivariate normal r.v. with zero mean and

a. for $i = 1, \ldots, (N - N_{out})$: covariance matrix $\mathbf{\Sigma}_1 = [a_{rs}]$, with $a_{rr} = 9$ and $a_{rs} = -1$ $(r \neq s)$, $r, s = 1, \ldots, 4$;

b. for $i = (N - N_{out}) + 1, \ldots, N$: covariance matrix $\mathbf{\Sigma}_{1.out} = [a_{rs}]$, with $a_{rr} = 25$ and $a_{rs} = 1$ $(r \neq s)$, $r, s = 1, \ldots, 4$; (outlier in the factor space);

- $(x_{i8}, x_{i9})^\top$ is distributed as a multivariate t-distribution with 3 degrees of freedom and scale matrix $\mathbf{\Sigma}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$;
- x_{i10} follows a Poisson distribution $\mathcal{P}(5)$.

Simulated response vectors

For each design matrix ${}_h\boldsymbol{X}$, we simulate S independent response vectors ${}_h\boldsymbol{Y}_s$, whose i-th item is

$$ {}_hY_{s,i} = {}_h\boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_{si}, \quad i = 1, \ldots, N, \quad \text{with} $$

- $\boldsymbol{\beta} = (1, 1, 1, 1, 2, 2, 2, 2, 1, 1, 1)$ and $\sigma = 3$ for $i = 1, \ldots, N - N_{out}$;
- $\boldsymbol{\beta} = (1, 1, 1, 1, -2, -2, -2, -2, 1, -1, -1)$, $\sigma = 20$ for $i = (N - N_{out}) + 1, \ldots, N$ (Outliers)

Different subsampling algorithms

At each simulation step (h, s), with $h = 1, \ldots, H$ and $s = 1, \ldots, S$, to draw a subsample $s_n^{(h,s)}$ from the simulated dataset ${}_h\boldsymbol{D}_s$, we have applied the following algorithms:

1. Non-informative I (Algorithm 4)
2. Non-informative D (Algorithm 2)
3. Informative I (Algorithm 4 with Cook's distance steps)
4. Informative D (Algorithm 3)
5. Simple random sampling (SRS): passive learning selection

To implement the I-optimality procedure, we have generated a **prediction set** $\mathcal{X}_0 = \{\boldsymbol{x}_{01}, \ldots, \boldsymbol{x}_{0N_0}\}$ without high leverage points ($N_0 = 500$).

Assessing the distinct subsampling methods

a) We check **goodness** of the subsampling methods wrt **D- and I-optimality criteria**.

b) To compare the performance of the different subsamples in terms of **prediction ability** on $\mathcal{X}_0$, we have also generated the corresponding responses (without outliers):

$$\boldsymbol{D}_0 = \{(\boldsymbol{x}_{01}, y_{01}), \ldots, (\boldsymbol{x}_{0N_0}, y_{0N_0})\}.$$

To **further** assess their **prediction ability** we have generated an independent **test set** $\mathcal{X}_T = \{\boldsymbol{x}_{T1}, \ldots, \boldsymbol{x}_{TN_T}\}$ of size $N_T = 500$ without high leverage points and the corresponding responses (without outliers):

$$\boldsymbol{D}_T = \{(\boldsymbol{x}_{T1}, y_{T1}), \ldots, (\boldsymbol{x}_{TN_T}, y_{TN_T})\}.$$

Monte Carlo averages for the subsamples of size $n = 500$

Optimality properties

Algorithm	$\text{MSPE}_{\mathcal{X}_0}$	Log(det)
Non-inf. I	**0.0857**	93.4269
Non-inf. D	0.0947	**94.3877**
Inf. I	0.0938	92.0869
Inf. D	0.1030	92.7748
SRS	0.2056	82.5234

- The **average mean squared prediction error** in $\mathcal{X}_0$:

$$\text{MSPE}_{\mathcal{X}_0}^{(h,s)} = \sigma^2 \frac{\text{Trace}\left[\left(\sum_{i=1}^{N} {}_h\boldsymbol{x}_i \, {}_h\boldsymbol{x}_i^\top I_\ell^{(h,s)}\right)^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0\right]}{N_0}, \quad I_\ell^{(h,s)} = \begin{cases} 1 & \text{if } \ell \in s_n^{(h,s)} \\ 0 & \text{otherwise} \end{cases}$$

- The **logarithm of the determinant of the precision matrix**:

$$\text{Log(det)}^{(h,s)} = \log \left| \sum_{i=1}^{N} {}_h\boldsymbol{x}_i \, {}_h\boldsymbol{x}_i^\top I_\ell^{(h,s)} \right|$$

Monte Carlo averages for the subsamples of size $n = 500$

Predictive abilities

Algorithm	$\mathrm{SPE}_{\mathcal{X}_0}$	$\mathrm{SPE}_{\mathcal{X}_T}$	$\mathrm{SE}_{\boldsymbol{D}_0}$	$\mathrm{SE}_{\boldsymbol{D}_T}$
Non-inf. I	6.5104	6.8020	16.0792	16.3538
Non-inf. D	6.1011	6.2945	15.5982	15.7969
Inf. I	**0.1464**	**0.1494**	**9.4445**	**9.5337**
Inf. D	0.1594	0.1601	9.4564	9.5448
SRS	0.2629	0.2671	9.5683	9.6594

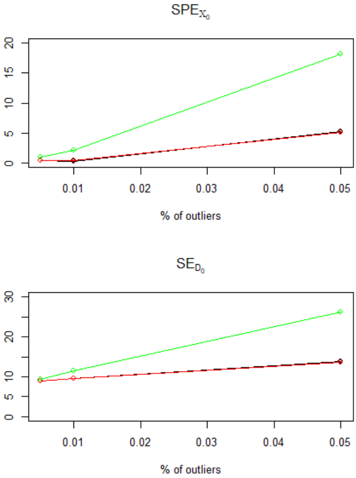
- The **average squared prediction error** in $\mathcal{X}_0$ and in $\mathcal{X}_T = \{\boldsymbol{x}_{T1}, \ldots, \boldsymbol{x}_{TN_T}\}$:

$$\mathrm{SPE}_{\mathcal{X}_0}^{(h,s)} = \frac{\sum_{i=1}^{N_0} (\hat{y}_{0i}^{(h,s)} - \mu_{0i})^2}{N_0} \quad \text{and} \quad \mathrm{SPE}_{\mathcal{X}_T}^{(h,s)} = \frac{\sum_{i=1}^{N_T} (\hat{y}_{Ti}^{(h,s)} - \mu_{Ti})^2}{N_T},$$

- The **standard error** in the prediction set $\boldsymbol{D}_0$ and in the test set $\boldsymbol{D}_T$:

$$\mathrm{SE}_{\boldsymbol{D}_0}^{(h,s)} = \frac{\sum_{i=1}^{N_0} \left(\hat{y}_{0i}^{(h,s)} - y_{0i}\right)^2}{N_0} \quad \text{and} \quad \mathrm{SE}_{\boldsymbol{D}_T}^{(h,s)} = \frac{\sum_{i=1}^{N_T} \left(\hat{y}_{Ti}^{(h,s)} - y_{Ti}\right)^2}{N_T}$$

Effect of a different percentage of outliers on predictions



SRS = green Inf.D = red Inf.I = black

7. Conclusion and future developments

Major features of our approach

- our approach may be implemented in a non-informative and informative setting, according to the available information
- it guarantees that the selected samples are the most informative for estimation (D-optimality) or predictive purposes (I-optimality) of the model generator of the majority of the data

Future developments

- Extension of the proposed algorithm to the generalized linear model
- Account for model uncertainty in presence of outliers
- Comparison with orthogonal and space-filling design subsampling methods (which are robust wrt misspecification model) in presence of outliers

References

- Atkinson, A., Donev, A., & Tobias, R. (2007). Optimum experimental designs, with SAS (Vol. 34). Oxford University Press.
- Chatterjee, S., & Hadi, A.S. (1986). Influential Observations, High Leverage Points, and Outliers in Linear Regression. Statistical Sciences
- Deldossi, L., & Tommasi, C. (2022). Optimal design subsampling from Big Datasets, Journal of Quality Technology
- Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. The American Statistician
- Wang, H., Yang, M., & Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. Journal of the American Statistical Association
- Yu, J., Ai, M., Ye, Z. (2023). A review on design inspired subsampling for big data, Statistical Papers

Thank you for your attention

6. A real data example

The diamonds data set in the `ggplot2` package contains the prices and the specifications for more than 50,000 diamonds. There are 7 factors in this data set:

1. the carat x_1, which is the weight of the diamond, ranges from 0.2 to 5.01
2. the quality x_2 of the diamond cut which is coded as one if the quality is better than "Very Good" and zero otherwise
3. the level of diamond color x_3 which is coded as one if the quality is better than "level F" and zero otherwise
4. a measurement of the diamond clearness x_4 wich takes value one if the quality is better than "SI1" and zero otherwise
5. the total depth percentage x_5
6. the width at the widest point x_6
7. the volume of the diamond x_7

The response variable y is log_{10} of the price. To avoid multicollinearity x_1 has not been considered (higly correlated with x_7).
Moreover the quadratic effect of x_7 has been included in the model.

GOAL

the prediction of the price of the diamonds with a volume larger than 200 mm^3

Cross-validation averages for the subsamples of size $n = 100$

- Prediction set $\mathcal{X}_0$ randomly selected from all the diamonds with x_7 larger than 200 mm^3.
- The remaining dataset has been divided in 4 folds of the same size. In rotation, one fold represents the test set, while the others form the training set
- In each test set only diamonds with volume larger than 200 mm^3 are considered and the outliers (if present) are removed.

Algorithm	$\mathrm{MSPE}_{\mathcal{X}_0}$	Log(det)	SE_{D_0}	SE_{D_T}
non-inf. I	**0.0452**	65.2964	0.0083	0.0092
non-inf. D	0.0602	**69.4402**	0.0569	0.0549
inf I	0.0454	65.1758	**0.0079**	**0.0084**
inf D	0.0620	65.9726	0.0097	0.0122
SRS	0.0998	60.9025	0.0117	0.0109

N.B. $\mathcal{X}_0$ and $\mathcal{X}_T$ include diamonds with a volume larger than 200 mm^3, $\tilde{N} = 2000$, $t_{max} = 2000$.

Algorithm 5: Initialization step for Algorithm 2 and 4

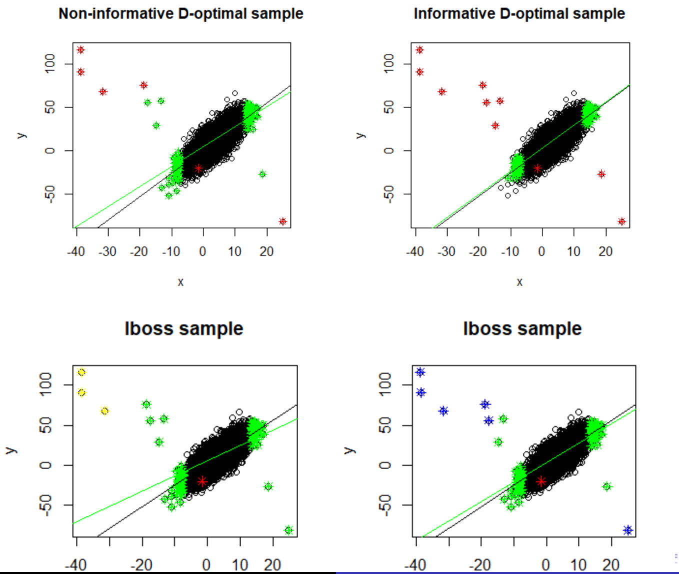
Goal: To find out an initial sample $s_n^{(0)}$ in the bulk of the data.

Algorithm 5 Initialization step for Algorithms 2 and 4

Require: Design matrix $\boldsymbol{X}$, sample size n, ν_2, t_{max}, $\tilde{N}$
Ensure: $s_n^{(0)}$: initial sample without high leverage points

From U select without replacement a simple random sample of size n, $r_n^{(0)}$
Set $t = 0$
while $t < t_{max}$ **do**
  Compute the leverage scores for the current sample
  $h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_i$, where $i \in r_n^{(t)}$
  Identify unit $i_m = \arg\max_{i \in r_n^{(t)}} h_{ii}$
  if $h_{i_m i_m} < \nu_2 \frac{k+1}{n}$ **then**
    Set $s_n^{(0)} = r_n^{(t)}$ and stop the iterative procedure
  else
    Select randomly $\tilde{N}$ units from $\left\{U - r_n^{(t)}\right\}$
    Let $\boldsymbol{x}_j$, with $j = 1, \dots, \tilde{N}$, the observations for these units
    Compute $({}_j\boldsymbol{X}_t^\top {}_j\boldsymbol{X}_t)^{-1}$ from (5), where ${}_j\mathbf{X}_t$ is the design matrix obtained from $\mathbf{X}_t$ exchanging $\mathbf{x}_{i_m}$ with $\mathbf{x}_j$
    Determine the leverage scores $h_{i_m i_m}(\boldsymbol{x}_j) = \boldsymbol{x}_j^\top ({}_j\boldsymbol{X}_t^\top {}_j\boldsymbol{X}_t)^{-1} \boldsymbol{x}_j$
    Identify the set of points candidate for the exchange with i_m:
    $\mathcal{C}^{(t)} = \left\{ j : \ h_{i_m i_m}(\boldsymbol{x}_j) < \nu_2 \frac{k+1}{n} \right\}$
    Select at random a unit j_a from $\mathcal{C}^{(t)}$
    Determine $r_n^{(t+1)}$ by replacing unit i_m with j_a in $r_n^{(t)}$
    Set $(\boldsymbol{X}_{t+1}^\top \boldsymbol{X}_{t+1})^{-1} = ({}_{j_a}\boldsymbol{X}_t^\top {}_{j_a}\boldsymbol{X}_t)^{-1}$
    Set $t = t + 1$
  end if
end while

Example (follows): Comparison with Iboss if outliers are removed after the sample selection (Nout=10)



Example (follows): Comparison with Iboss if outliers are removed after the sample selection (Nout=200)