Statistical Papers (2023) 64:1119–1135
https://doi.org/10.1007/s00362-023-01422-3

REGULAR ARTICLE

Accounting for outliers in optimal subsampling methods

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

Received: 12 December 2022 / Revised: 27 February 2023 / Published online: 27 April 2023

Abstract
Nowadays, in many different fields, massive data are available and for several reasons, it might be convenient to analyze just a subset of the data. The application of the D-optimality criterion can be helpful to optimally select a subsample of observations. However, it is well known that D-optimal support points lie on the boundary of the design space and if they go hand in hand with extreme response values, they can have a severe influence on the estimated linear model (leverage points with high influence). To overcome this problem, firstly, we propose a non-informative "exchange" procedure that enables us to select a "nearly" D-optimal subset of observations without high leverage values. Then, we provide an informative version of this exchange procedure, where besides high leverage points also the outliers in the responses (that are not necessarily associated to high leverage points) are avoided. This is possible because, unlike other design situations, in subsampling from big datasets the response values may be available. Finally, both the non-informative and informative selection procedures are adapted to I-optimality, with the goal of getting accurate predictions.

Keywords D-optimality · I-optimality · Active learning · Subsampling

Mathematics Subject Classification 62K05 · 62D99 · 62F35 · 62J05

Elena Pesce and Chiara Tommasi have contributed equally to this work.

✉ Laura Deldossi
laura.deldossi@unicatt.it

Elena Pesce
elena_pesce@swissre.com

Chiara Tommasi
chiara.tommasi@unimi.it

[1] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, L.go Gemelli 1, 20123 Milan, Italy

[2] Swiss Re Institute, Swiss Re Management Ltd, Mythenquai 50/60, 8022 Zurich, Switzerland

[3] Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, 20122 Milan, Italy

1 Introduction

Recently, the theory of optimal design has been exploited to draw a subsample from huge datasets, containing the most information for the inferential goal; see Drovandi et al. (2017), Wang et al. (2019), Deldossi and Tommasi (2022) among others. Unfortunately, Big Data sets usually are the result of passive observations, so some high leverage values in the covariates and/or outliers in the response variable (denoted by Y) may be present. In this study, we assume that a small percentage of the data are outliers and the goal is to provide a precise estimate of the model parameters or an accurate prediction for the model that generates the majority of the data.

The most commonly applied criterion is the D-optimality. It is well known that D-optimal designs tend to lie on the boundary of the design region thus, in the presence of high leverage values, all of them would be selected. Since this circumstance could have a severe influence on the estimated model (leverage points with high influence), we propose an "exchange" procedure to select a "nearly" D-optimal subset which does not include high leverage values. Avoiding high leverage points, however, does not guard from all the outliers in Y. Therefore, we also modify the previous method to exploit the information about the responses and avoid the selection of the abnormal Y-values. The first proposal is a non-informative procedure, as it is not based on the response observations, while the latter is an informative exchange method.

Finally, both these exchange algorithms are adapted to the I-criterion, which aims at providing accurate predictions in a set of covariate-values (called prediction set).

Notation and motivation of the work are introduced in Sect. 2. Section 3 describes the novel modified exchange algorithm to obtain both non-informative and informative D-optimal subsamples without outliers. In Sect. 4 we adapt our proposal to I-optimality, to select a subsample with the goal of obtaining accurate predictions. In Sect. 5 we develop some simulations and a real data example, to assess the performance of the proposed subsampling methods. Finally, in Appendix we suggest a procedure for the initialization of these algorithms.

2 Notation and motivation of the work

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

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

where $^\top$ denotes transposition, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ is a vector of unknown coefficients, $\boldsymbol{x}_i^\top = (1, \tilde{\boldsymbol{x}}_i^\top)$ where $\tilde{\boldsymbol{x}}_i = (x_{i1}, \ldots, x_{ik})^\top$, for $i = 1, \ldots, N$, are N iid repetitions of a k-variate explanatory variable, and ε_i are iid random errors with zero mean and equal variance σ^2.

$\boldsymbol{D} = \{(\tilde{\boldsymbol{x}}_1^\top, Y_1), \ldots, (\tilde{\boldsymbol{x}}_N^\top, Y_N)\}$ indicates the available dataset, which is assumed to be a tall dataset, i.e. with $k << N$.

The population under study is denoted by $U = \{1, \ldots, N\}$ and $s_n \subseteq U$ denotes a sample without replications of size n from U (i.e. a collection of n different indices from U).

Herein, we describe a new sampling method from a given dataset $\boldsymbol{D}$, with the goal of selecting n observations ($k < n << N$) to produce an efficient parameter estimate or an accurate prediction for the model generating the whole dataset apart from a few outliers, i.e. a small quantity of points that take "abnormal" values with respect to the rest of the data and that possibly have been generated by a different model.

Given a sample $s_n = \{i_1, \ldots, i_n\}$, let $\boldsymbol{X}$ be the $n \times (k+1)$ matrix whose rows are $\boldsymbol{x}_i^\top$, for $i \in s_n$, and let $\boldsymbol{Y} = (Y_{i_1}, \ldots, Y_{i_n})^\top$ be the $n \times 1$ vector of the sampled responses. We consider the OLS estimator of the coefficients of the linear model based on the sample s_n:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} \\ &= \left(\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T I_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i Y_i I_i,\end{aligned}$$

where

$$I_i = \begin{cases} 1 & \text{if } i \in s_n \\ 0 & \text{otherwise} \end{cases}, \qquad \text{with } i = 1, \ldots, N \tag{1}$$

denotes the sample inclusion indicator.

To improve the precision of $\hat{\boldsymbol{\beta}}$, we suggest to select the sample s_n according to D-optimality. We denote the D-optimum sample as

$$s_n^* = \underset{s_n = \{I_1, \ldots, I_N\}}{\arg\sup} \left| \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top I_i \right| .$$

When $\boldsymbol{D}$ contains outliers, s_n^* may include them, because D-optimal support points usually lie on the boundary of the experimental region. Example 1 illustrates this issue.

Example 1 An artificial dataset $\boldsymbol{D}$ with $N = 10000$ observations has been generated from a simple linear model,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, N,$$

in the following way:
for $i = 1, \ldots, 9990$, $\boldsymbol{\beta} = (1.5, 2.7)^\top$, $x_i \sim \mathcal{N}(3, 4)$, $\varepsilon_i \sim \mathcal{N}(0, 9^2)$;
for $i = 9991, \ldots, 10000$, $\boldsymbol{\beta} = (1.5, -2.7)^\top$ $x_i \sim \mathcal{N}(3, 20)$, $\varepsilon_i \sim \mathcal{N}(0, 20^2)$.
The left-hand side of Fig. 1 displays these last 10 observations which are isolated with respect to the majority of the data, generated from the first distribution. The right-hand side of Fig. 1 emphasises the D-optimal subsample of size $n = 100$, s_n^*. As expected, all the abnormal values in X are included in s_n^* because they maximize the determinant of the information matrix [s_n^* has been obtained by applying the function `od_KL` of the `R` package `OptimalDesign` (Harman and Filová, 2019)].

A similar behaviour would be displayed also by the I-optimal subsample, that should be applied to get accurate predictions (see Sect. 4). To avoid the inclusion of outliers

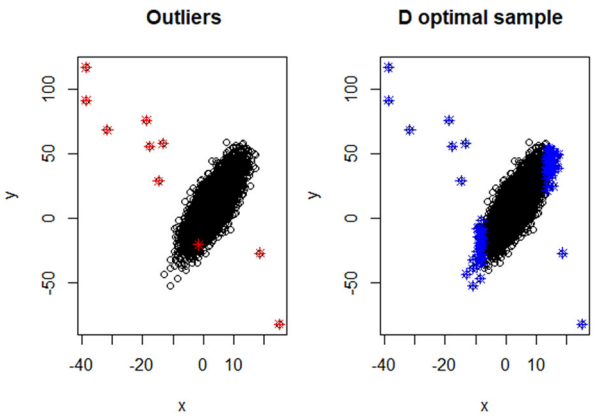

Fig. 1 Outliers (in red) and the D-optimal sample (in blue). (Color figure online)

when applying the D- or I-optimal subsampling, we propose a modification of the well known exchange algorithm.

Before describing our proposal, we recall that a tool to identify an outlier in the factor-space is the leverage score, $h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i$. Points which are isolated in the factor-space (i.e., far located from the main body of points) can be thought of as outliers and are characterized by high leverage values [see Chatterjee and Hadi (1986)]. Actually, an observation $\boldsymbol{x}_i$, with $i = 1, \cdots, n$, such that

$$h_{ii} > \nu_1 \frac{k+1}{n},$$

where ν_1 is a tuning parameter usually set equal to 2 [see for instance Hoaglin and Welsch (1978)], that is called a *high leverage point*. In general, high leverage points allow to reduce the variance of the parameters' estimates and in the literature many leverage-based sampling procedure have been proposed [see, among others Ma et al. (2015)]. But consider that if these high leverage points are associated to outlying response values, their inclusion in the sample may lead to misleading inferential results. For this reason, our aim is to avoid these points.

3 Modified exchange algorithms

The common structure of the t-th iteration of an exchange algorithm consists in adding a unit, chosen from a list of candidate points $\mathcal{C}^{(t)}$, to the current sample $s_n^{(t)}$, and then deleting an observation from it. The choice of the augmented and deleted points is based on the achievement of some optimality criterion. For instance, for D-optimality, Algorithm 1 describes the classical exchange procedure [see Chapter 12 in Atkinson et al. (2007)].

Our main idea is to modify Algorithm 1 by not proposing for the exchange the high leverage points, thus avoiding the inclusion in the sample of high leverage scores with

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 $j_a = \underset{j \in \mathcal{C}^{(t)}}{\arg\max}\, \boldsymbol{x}_j^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j$
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 $i_m = \underset{i \in s_{n+1}^{(t)}}{\arg\min}\, h_{ii}$
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**

abnormal responses, which could lead to wrong inferential conclusions. This goal is reached by:

(a) Switching the augmentation and deletion steps;
(b) Changing the set $\mathcal{C}^{(t)}$ where the observation to be added is searched.

If the information about the responses is not exploited in step b (to identify $\mathcal{C}^{(t)}$), then the modified D-optimal sample is non-informative for the parameters of interest. The non-informative procedure is described in detail in Subsect. 3.1.

Preventing high leverage points, however, does not guard from all the outliers in Y: there may exist points that are in the core of the data with respect to the features, while being abnormal with respect to the response variable. In Subsect. 3.2 we propose another version of the algorithm, where (in step b) we employ the responses to remove the outliers in Y. Note that the obtained optimal subsample becomes informative because of the dependence on the Y values.

3.1 Non-informative D-optimal samples without high leverage points

Let $s_n^{(t)}$ be the current sample of size n and $s_n^{(0)}$ an initial sample which does not include high leverage points (see Algorithm 5 in Appendix for a detailed procedure to get a convenient initial sample).

To update $s_n^{(t)}$, firstly we remove from it the unit i_m with the smallest leverage score,

$$i_m = \underset{i \in s_n^{(t)}}{\arg\min}\, \boldsymbol{x}_i^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_i = \underset{i \in s_n^{(t)}}{\arg\min}\, h_{ii},$$

thus obtaining a reduced sample of size $n-1$, where $\boldsymbol{X}_t$ denotes the design matrix associated to $s_n^{(t)}$.

Let $\boldsymbol{X}_t^-$ be the design matrix attained by leaving out the row $\boldsymbol{x}_{i_m}$ from $\boldsymbol{X}_t$. Subsequently, we add the unit $j_a \in \mathcal{C}^{(t)}$ with the largest leverage score $\boldsymbol{x}_{j_a}^\top (\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^-)^{-1} \boldsymbol{x}_{j_a}$,

where the set of candidate points for the exchange at the current iteration is

$$\mathcal{C}^{(t)} = \left\{ j : \; h_{i_m i_m} < h_{i_m i_m}(\boldsymbol{x}_j) < \nu_1 \frac{k+1}{n} \right\}, \tag{2}$$

$$(\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^{-})^{-1} = (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} + (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \frac{\boldsymbol{x}_{i_m} \boldsymbol{x}_{i_m}^T}{1 - \boldsymbol{x}_{i_m}^T (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_{i_m}} (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1}, \tag{3}$$

[see Searle (1982) p. 153 to get (3)] and $h_{i_m i_m}(\boldsymbol{x}_j)$ is the leverage score obtained by exchanging $\boldsymbol{x}_{i_m}$ with $\boldsymbol{x}_j$ for $j \in \{U - s_n^{(t)}\}$. The next theorem provides an analytical expression for $h_{i_m i_m}(\boldsymbol{x}_j)$, which reduces the computational burden of the algorithm.

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*

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

where

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

with

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j \, (\boldsymbol{x}_j \boldsymbol{x}_{i_m}^\top + \boldsymbol{x}_{i_m} \boldsymbol{x}_j^\top) + [1 - \boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_{i_m}] \, \boldsymbol{x}_j \boldsymbol{x}_j^\top \\ &\quad - [1 + \boldsymbol{x}_j^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j] \, \boldsymbol{x}_{i_m} \boldsymbol{x}_{i_m}^\top; \\ d &= [1 - \boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_{i_m}] \, [1 + \boldsymbol{x}_j^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j] + [\boldsymbol{x}_{i_m}^\top (\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \boldsymbol{x}_j]^2. \end{aligned}$$

Proof Expression (5) can be obtained from Lemma 3.3.1 in Fedorov (1972) after some cumbersome algebra. □

In force of the upper bound in (2), our proposal is to consider as candidates for the exchange only observations in $\{U - s_n^{(t)}\}$ which are not high leverage points. In addition, to speed up the algorithm we reduce the number of exchanges by imposing the lower bound in (2). Without this lower bound, if $h_{i_m i_m}(\boldsymbol{x}_j) \leq h_{i_m i_m}$, the new observation j could be removed at the subsequent iteration.

Algorithm 2 outlines the steps to select a D-optimal subsample without high leverage points.

3.2 Informative D-optimal sample without outliers

Whenever the response values are available, this information should be exploited by the exchange algorithm, obtaining an informative D-optimal subsample.

According to Chatterjee and Hadi (1986) an influential data point in Y is an observation that strongly influences the fitted values. To identify these influential values, we adopt Cook's distance, but other measures can be similarly applied. Cook's distance

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 $\mathcal{C}^{(t)}$ 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: 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**

for the i-th observation, C_i, quantifies how much all of the fitted values in the model change when the i-th data point is deleted:

$$C_i = \frac{(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)})^{\top} (\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)})}{(k+1)\hat{\sigma}^2} = \frac{(Y_i - \hat{Y}_i)^2}{(k+1)\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}, \quad i = 1, \ldots, n, \tag{6}$$

where $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\top}$, $\hat{\sigma}^2$ is the residual mean square estimate of σ^2 and $\hat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{(i)}^{\top}$ is the vector of predicted values when the i-th unit is removed from the data set $\boldsymbol{D}$. According to a general practical rule, any observation with a Cook's distance larger than $4/n$ may be considered as an influential point.

To get an informative D-optimal sample, Algorithm 2 is modified by including the additional steps illustrated in Algorithm 3.

Algorithm 3 Informative optimal subsample without outliers: additional steps to be included between 7 and 8 in Algorithm 2 (and Algorithm 4)

Require: Dataset $\boldsymbol{D}$, sample size n
Ensure: Informative D-optimal sample without outliers
1: Compute Cook's distance for unit j_a:

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

2: **if** $C_{j_a} < 4/n$ **then**
3: go ahead to step 8 of Algorithm 2 (and Algorithm 4)
4: **else**
5: reject the exchange and go back to step 5 of Algorithm 2 (and Algorithm 4)
6: **end if**

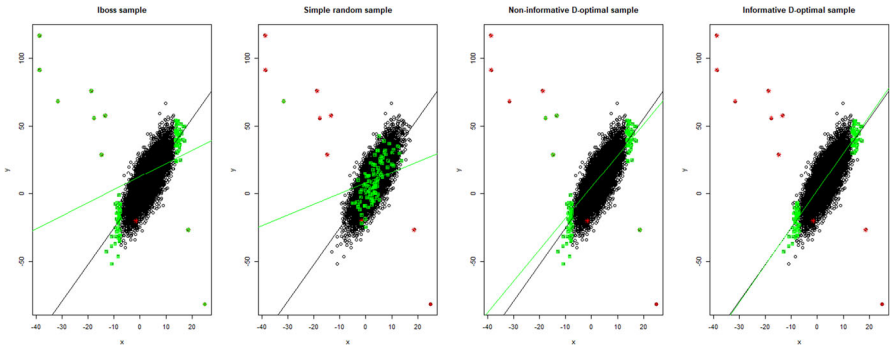

Fig. 2 Subsamples (in green) of the artificial dataset of Example 1, obtained applying: Iboss, Simple Random Sampling, Non-informative D-optimal sampling, Informative D-optimal sampling. The black line is the true regression model, while the green line is the fitted model based on the subsample (of size $n = 100$) selected according to the different procedures. (Color figure online)

Example 2 Figure 2 illustrates the performance of the proposed algorithms in comparison with the Iboss subsampling method [proposed by Wang et al. (2019)] and the simple random sample, in the artificial dataset of Example 1.

As expected, the Iboss algorithm provides a subset similar to the D-optimal sample (cfr. with Fig. 1) since it selects the points on the boundary of the design space, thus including most of the outliers. As a consequence, the true model and the fitted model are quite distinct. Neither the simple random sample produces a good filted model, as it includes an outlier. The non-informative selection procedure seems to improve the fit of the true model, even if the best performance is obtained using the informative selection approach, which doesn't include outliers.

Remark Let us note that an increase of t_{max} and $\tilde{N}$ would lead to an improvement of the D-optimal subsamples, because of a better chance of exchanging sample points. In particular, it is reasonable to consider $\tilde{N} = N - n$ whenever N is not too large.

4 Optimal subsampling to get accurate predictions

If we are interested in obtaining accurate predictions on a set of values $\mathcal{X}_0 = \{\boldsymbol{x}_{01}, \ldots, \boldsymbol{x}_{0N_0}\}$ instead of a precise parameter estimation, then we should select the observations minimizing the overall prediction variance. Let $\hat{Y}_{0i} = \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}_{0i}$ be the prediction of $\mu_{0i} = E(Y_{0i}|\boldsymbol{x}_{0i})$ at $\boldsymbol{x}_{0i}$, $i = 1, \ldots, N_0$. The prediction variance at $\boldsymbol{x}_{0i}$, also known as "mean squared prediction error" is

$$\text{MSPE}(\hat{Y}_{0i}|\boldsymbol{x}_{0i}, \boldsymbol{X}) = \text{E}[(\hat{Y}_{0i} - \mu_{0i})^2|\boldsymbol{x}_{0i}, \boldsymbol{X}].$$

If $\boldsymbol{X}_0$ is the $N_0 \times k$ matrix whose i-th row is $\boldsymbol{x}_{0i}^\top$, then a measure of the overall mean squared prediction error is the sum of the prediction variances in $\mathcal{X}_0$:

$$\sum_{i=1}^{N_0} \text{MSPE}(\hat{Y}_{0i}|\boldsymbol{x}_{0i}, \boldsymbol{X}) = \sigma^2 \,\text{trace}[\boldsymbol{X}_0 \,(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_0^\top]. \tag{7}$$

The following sample

$$s_n^I = \underset{s_n=\{I_1,\ldots,I_N\}}{\arg\inf} \;\text{trace}\left[\left(\sum_{i=1}^{N} \boldsymbol{x}_i \,\boldsymbol{x}_i^\top I_i\right)^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0\right]$$

minimizes the overall prediction variance (7) and is called I-optimal. It is well known that to produce accurate predictions it would be advisable to avoid outliers. An I-optimal subsample without high leverage points can be obtained by modifying the deletion and augmentation steps of the exchange algorithm described in Sect. 3.1 accordingly to the I-criterion. The current sample $s_n^{(t)}$ should be updated by removing the unit i_m which minimises the increase in the overall mean squared prediction error. From the results given in Appendix A of Meyer and Nachtsheim (1995), the increment in the overall mean squared prediction error due to the omission of the unit i is given by

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

where $\boldsymbol{X}_t$ is the $n \times k$ matrix whose rows are $\boldsymbol{x}_i^T$ with $i \in s_n^{(t)}$.

Subsequently, to obtain again a sample of size n, from a set $\mathcal{C}^{(t)}$ of candidate points, we should add the unit j_a which maximize the decrease in the overall mean squared prediction error:

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

where $\boldsymbol{X}_t^-$ is the design matrix obtained by removing the row $\boldsymbol{x}_{i_m}$ from $\boldsymbol{X}_t$ and $(\boldsymbol{X}_t^{-\top} \boldsymbol{X}_t^-)^{-1}$ can be computed from (3).

The set of candidate points should be composed by units that are not at risk to be deleted at the next iteration and are not high leverage points:

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

where $h_{i_m i_m}(\boldsymbol{x}_j)$ is given in (4),

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

${}_j\boldsymbol{X}_t$ 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 Eq. (5).

Algorithm 4 summarizes the steps to select a non-informative I-optimal sample, while to obtain its informative version, it is enough to incorporate the additional steps of Algorithm 3.

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

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

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

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

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**

5 Numerical studies

5.1 Simulation results

In this section, we evaluate the performance of our proposals through a simulation study. We generate $H \times S$ random datasets of size $N = 10^6$, each one including $N_{out} = 500$ high leverage points/outliers (with $H = 30$ and $S = 50$). The computation of some metrics will illustrate the validity of our procedures in selecting D- or I-optimal subsamples without outliers.

Precisely, for each $h = 1, \ldots, H$, N iid repetitions of a 10-variate explanatory variable ${}_h\tilde{\boldsymbol{x}}_i = (x_{i1}, \ldots, x_{i10})^\top$ are generated as follows:

1. x_{i1}, x_{i2} and x_{i3}, for $i = 1, \ldots, N$, are independently distributed as $U(0, 5)$;
2. $(x_{i4}, x_{i5}, x_{i6}, x_{i7})^\top$ is distributed as a multivariate normal r.v. with zero mean and
   - 2.a. For $i = 1, \ldots, (N - N_{out})$: covariance matrix $\boldsymbol{\Sigma}_1 = [a_{rs}]$, with $a_{rr} = 9$ and $a_{rs} = -1$ $(r \neq s)$, $r, s = 1, \ldots, 4$;
   - 2.b. For $i = (N - N_{out}) + 1, \ldots, N$: covariance matrix $\boldsymbol{\Sigma}_{1.out} = [a_{rs}]$, with $a_{rr} = 25$ and $a_{rs} = 1$ $(r \neq s)$, $r, s = 1, \ldots, 4$;

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

For each generated $N \times (k+1)$ design matrix ${}_h\boldsymbol{X}$, whose i-th row is ${}_h\boldsymbol{x}_i^\top = (1, {}_h\tilde{\boldsymbol{x}}_i^\top)$ $(i = 1, \ldots, N)$, we have simulated $S = 50$ independent $N \times 1$ response vectors ${}_h\boldsymbol{Y}_s$ (with $s = 1, \ldots, S$), whose i-th item is

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

with

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

At each simulation step (h, s), to draw subsamples from the simulated dataset:

$$ {}_h\boldsymbol{D}_s = \{({}_h\boldsymbol{x}_1^\top, {}_hy_{s,1}), \ldots, ({}_h\boldsymbol{x}_N^\top, {}_hy_{s,N})\}, $$

we have applied the following algorithms ($h = 1, \ldots, H$ and $s = 1, \ldots, S$):

1. Non-informative I (Algorithm 4)
2. Non-informative D (Algorithm 2)
3. Informative I (Algorithms 4 and 3)
4. Informative D (Algorithms 2 and 3)
5. Simple random sampling (SRS): passive learning selection

To assess these subsampling techniques, we have generated a test set of size $N_T = 500$, without high leverage points and outliers (i.e. with $N_{out} = 0$):

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

Finally, to implement the I-optimality procedure, we have generated a prediction region $\mathcal{X}_0$ without high leverage points. In addition, to compare the performance of the distinct subsamples in terms of prediction ability on $\mathcal{X}_0$, we have generated also the corresponding responses (without outliers). Let

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

be the prediction set, where $N_0 = 500$.

A subsample selected from the dataset ${}_h\boldsymbol{D}_s$ (generated at the (h, s)-th simulation step) is denoted by $s_n^{(h,s)}$, and

$$ I_i^{(h,s)} = \begin{cases} 1 & \text{if } i \in s_n^{(h,s)} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots, N, \tag{10} $$

is the corresponding sampling indicator variable, for $h = 1, \ldots, H$ and $s = 1, \ldots, S$.

At each simulation step (h, s):

Table 1 Monte Carlo averages $\text{MSPE}_{\mathcal{X}_0}$ and Log(det) for the subsamples of size 500 obtained from the different algorithms (in bold the best values assumed by the Monte Carlo averages and Log(det))

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

Bold values assumed by the Monte Carlo averages ande Log(det)

(a) To evaluate the performance of the subsampling techniques with respect to D- and I-optimality criteria, we have computed:
– The average mean squared prediction error in $\mathcal{X}_0$ [from (7)]:

$$\text{MSPE}_{\mathcal{X}_0}^{(h,s)} = \sigma^2 \frac{\text{trace}\left[\left(\sum_{i=1}^{N} {}_h\boldsymbol{x}_i \, {}_h\boldsymbol{x}_i^\top I_i^{(h,s)}\right)^{-1} \boldsymbol{X}_0^\top \boldsymbol{X}_0\right]}{N_0};$$

– The logarithm of the determinant of the information matrix:

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

(b) To assess the predictive ability of the selection algorithms, we have considered:
– The average squared prediction error in $\mathcal{X}_0$ and in $\mathcal{X}_T = \{\boldsymbol{x}_{T1}, \ldots, \boldsymbol{x}_{TN_T}\}$:

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

where $\hat{y}_{0i}^{(h,s)} = {}_h\hat{\boldsymbol{\beta}}_s^\top \boldsymbol{x}_{0i}$, $\hat{y}_{Ti}^{(h,s)} = {}_h\hat{\boldsymbol{\beta}}_s^\top \boldsymbol{x}_{Ti}$, $\mu_{0i} = \boldsymbol{\beta}^\top \boldsymbol{x}_{0i}$, $\mu_{Ti} = \boldsymbol{\beta}^\top \boldsymbol{x}_{Ti}$ and ${}_h\hat{\boldsymbol{\beta}}_s$ is the OLS estimate of $\boldsymbol{\beta}$ based on the subsample $s_n^{(h,s)}$;
– The standard error in the prediction set $\boldsymbol{D}_0$ and in the test set $\boldsymbol{D}_T$:

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

Table 1 displays the following Monte Carlo averages,

$$\text{MSPE}_{\mathcal{X}_0} = \frac{\sum_{h=1}^{H} \sum_{s=1}^{S} \text{MSPE}_{\mathcal{X}_0}^{(h,s)}}{HS} \quad \text{and} \quad \text{Log(det)} = \frac{\sum_{h=1}^{H} \sum_{s=1}^{S} \text{Log(det)}^{(h,s)}}{HS},$$

for the different sampling strategies: non-inf. I, non-inf. D, inf. I, inf. D and SRS, respectively. The results have been obtained having set $n = 500$, $\tilde{N} = 1000$, $t_{max} = 500$, $\nu_1 = 2$ and $\nu_2 = 3$.

From Table 1, the non-informative procedures seem to provide subsamples "nearly" D- and I-optimal that do not include high leverage points (they would be exactly D- and I-optimal if they allowed for these abnormal values). This result is consistent with the definitions of I- and D-optimality.

Table 2 instead lists the following Monte Carlo averages:
$\mathrm{SPE}_{\mathcal{X}_0} = \sum_{h=1}^{H}\sum_{s=1}^{S}\mathrm{SPE}_{\mathcal{X}_0}^{(h,s)}/HS$, $\mathrm{SPE}_{\mathcal{X}_T} = \sum_{h=1}^{H}\sum_{s=1}^{S}\mathrm{SPE}_{\mathcal{X}_T}^{(h,s)}/HS$,
$\mathrm{SE}_{D_0} = \sum_{h=1}^{H}\sum_{s=1}^{S}\mathrm{SE}_{D_0}^{(h,s)}/HS$ and $\mathrm{SE}_{D_T} = \sum_{h=1}^{H}\sum_{s=1}^{S}\mathrm{SE}_{D_T}^{(h,s)}/HS$,
for the different subsamples. These quantities enable to assess the predictive ability of the subsampling techniques. From Table 2, we can appreciate the prominent role of the informative procedures. In fact, when the database includes outliers in Y which are not associated with high leverage points (as in this simulation study), only the informative procedures are able to exclude them providing accurate predictions.

From the last row of Table 2, the SRS seems to behave quite well: it is fast, easy to be implemented and provides good predictions compared to the informative I-optimal subsampling. However, such a nice performance is due to the low percentage of outliers present in the artificial datasets. Figure 3 displays the superiority of the informative procedures with respect to the passive learning selection (SRS), as the percentage of the outliers increases. Of course, we consider a short range for the percentage of outliers because outliers are (by definition) a few isolated data points.

Comparing the third and the fourth rows of Table 2, informative I-optimal subsamples seem outperform the D-optimal ones only slightly, despite I-optimality should reflect the goal of getting accurate predictions. This happens because the prediction set $\mathcal{X}_0$ has a similar shape as the dataset. When $\mathcal{X}_0$ defines a specific subset of covariate-values, then the superiority of I-optimality emerges. See for instance the values of $\mathrm{SPE}_{\mathcal{X}_0}$ and $\mathrm{SPE}_{\mathcal{X}_T}$ in Table 3, where $\mathcal{X}_0$ and $\mathcal{X}_T$ involve only positive values of the features.

Table 2 Monte Carlo averages $\mathrm{SPE}_{\mathcal{X}_0}$, $\mathrm{SPE}_{\mathcal{X}_T}$, SE_{D_0} and SE_{D_T} for the subsamples of size 500 obtained from the different algorithms (in bold the minimal values assumed by the Monte Carlo averages)

Algorithm	$\mathrm{SPE}_{\mathcal{X}_0}$	$\mathrm{SPE}_{\mathcal{X}_T}$	SE_{D_0}	SE_{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

Bold minimal values assumed by the Monte Carlo averages

Table 3 Monte Carlo averages $\mathrm{SPE}_{\mathcal{X}_0}$, $\mathrm{SPE}_{\mathcal{X}_T}$, SE_{D_0} and SE_{D_T} for the subsamples of size 500 obtained from the different algorithms, when $\mathcal{X}_0$ and $\mathcal{X}_T$ are subsets of positive values (in bold the minimal values assumed by the Monte Carlo averages)

Algorithm	$\mathrm{SPE}_{\mathcal{X}_0}$	$\mathrm{SPE}_{\mathcal{X}_T}$	SE_{D_0}	SE_{D_T}
Non-inf. I	5.2506	5.1314	14.7100	13.9918
Non-inf. D	14.8542	14.6276	24.3856	23.5669
Inf. I	**0.1038**	**0.1084**	**9.3807**	**9.2020**
Inf. D	0.1544	0.1546	9.4306	9.2514
SRS	0.3150	0.3256	9.5963	9.4073

Bold minimal values assumed by the Monte Carlo averages

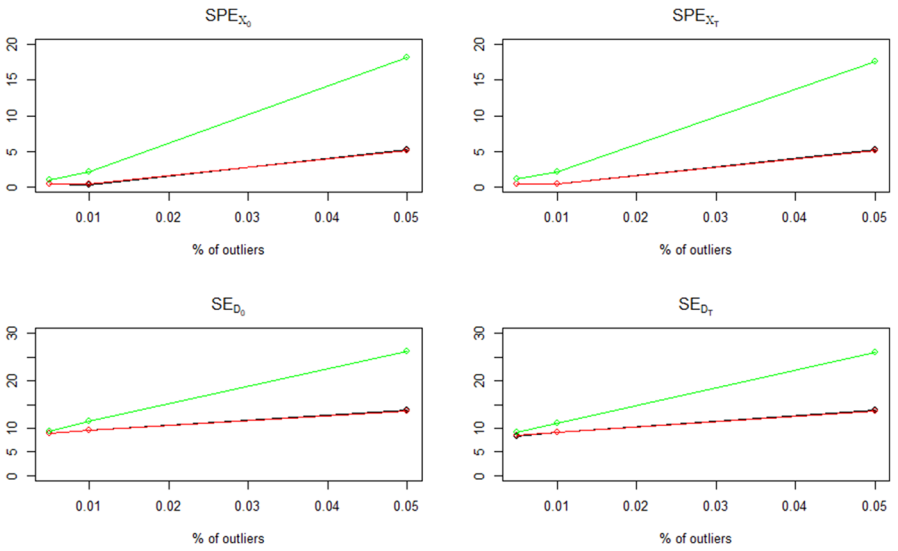

Fig. 3 Monte Carlo averages $\text{SPE}_{\mathcal{X}_0}$, $\text{SPE}_{\mathcal{X}_T}$, SE_{D_0} and SE_{D_T} for subsamples of size 500, including a different percentage of outliers (inf. I = black, inf. D = red, SRS = green). (Color figure online)

Remark Actually, to take into account the randomness of the SRS technique, we have drawn $N_{SRS} = 50$ different independent SRSs from each dataset ${}_h\boldsymbol{D}_s$, for $h = 1, \ldots, H$ and $s = 1, \ldots, S$; the Monte Carlo averages for SRS are based also on these additional observations.

5.2 Real data example

In this section we apply our proposal to the diamonds data set in the `ggplot2` package. This dataset contains the prices and the specifications for more than 50000 diamonds. More specifically, 7 features are included:

- The carat x_1, which is the weight of the diamond and ranges from 0.2 to 5.01;
- The quality of the diamond cut x_2, which is coded by one if the quality is better than "Very Good" and zero otherwise;
- The level of diamond color x_3, which is coded by one if the quality is better than "level F" and zero otherwise;
- A measurement of the diamond clearness x_4, which takes value one if the quality is better than "SI1" and zero otherwise;
- The total depth percentage x_5;
- The width at the widest point x_6;
- The volume of the diamond x_7.

To avoid a multicollinearity problem, x_1 has not been considered in the analysis, because it is highly correlated with x_7 (the volume). Furthermore, to obtain a better fit of the data, the quadratic effect of x_7 has been included in the model, where the response variable Y is the logarithm of the price ($\log_{10}$).

Table 4 Cross-validation averages of $\text{MSPE}_{\mathcal{X}_0}$, Log(det), SE_{D_0} and SE_{D_T} for the subsamples of size 100 obtained from the different algorithms, when $\mathcal{X}_0$ and $\mathcal{X}_T$ include diamonds with a volume larger than 200 mm^3 (in bold the best values assumed by the Cross-validation averages)

Algorithm	$\text{MSPE}_{\mathcal{X}_0}$	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

The dataset contains some outliers, such as observation NO.24068 which corresponds to a diamond with an unusually large width that makes the price too high.

Let us assume that the goal is the prediction of the price of the diamonds with a volume larger than 200 mm^3. Therefore, to apply the I-optimality strategy, we have randomly selected a prediction set $\mathcal{X}_0$ from all the diamonds with x_7 larger than 200 mm^3. Then, the remaining dataset has been divided in fourfolds of the same size to compare the different subsampling techniques through a cross-validation approach. In rotation, one fold represents the test set, while the others form the training set, from which subsample of size n are selected according to the different algorithms. In each test set only diamonds with volume larger than 200 mm^3 are considered; in addition, the outliers (if present) are removed. In this example we have set $n = 100$, $\tilde{N} = 2000$, $t_{max} = 2000$.

The first two columns of Table 4 show that the minimum value of the $\text{MSPE}_{\mathcal{X}_0}$ is associated to the non-informative I-Algorithm, while the maximum value of Log(Det) corresponds to the non-informative D-optimal subsample. This result is consistent with the definitions of I- and D-optimality and with the results of the simulation study in Sect. 5. With regards to the predictive ability of the subsampling techniques, we can observe that the I-informative procedure leads to the minimum values of the Cross-validation averages SE_{D_0} and SE_{D_T} (last two columns of Table 4). Differently from the simulation study, in this real data example, also the non-informative I-criterion seems to perform properly. This is due to the fact that in the diamonds dataset most of the outliers in Y are associated with high leverage points and thus also the non-informative procedure is able to exclude them providing accurate predictions.

6 Discussion

Recent advances in technology have brought the ability to collect, transfer and store large datasets. The availability of such a huge amount of data is a great challenge nowadays. However, very often Big Datasets contain noisy data because they are the result of a passive observation and not of a well planned survey. Moreover, huge datasets may not be queried for free; typically agencies that create and manage huge databases, enable to download data by paying a price per MB. Furthermore, there are circumstances where the value of the response variable may be obtained only for a restricted number of units.

For this reason, we suggest to consider only a subsample of the dataset excluding abnormal values, with the idea that a subset of a few relevant data may be more "informative" than a huge quantity of raw, redundant, and noisy observations. The theory of optimal design is a guide to draw a subsample containing the most informative observations, but optimal subsamples frequently lie on the boundary of the factor-domain, including all the outliers. Two modifications of the well-known exchange algorithm are herein proposed to select "nearly" optimal subsamples without abnormal values:

- A non-informative procedure, that avoids the inclusion of high leverage points, can be applied whenever information about the responses is not available or is too expensive to have it;
- An informative procedure, that excludes outliers in the response besides high leverage points, can be used whenever the responses are available.

A simulation study confirms that D-optimal subsampling should be applied if the inferential goal is precise estimation of the parameters, while informative I-optimal algorithm should be applied to get accurate predictions on a specified prediction set.

A limitation of these methods is that they are model-based, while in the real-life problems the model is unknown. This relevant issue will be handled in a future research by adapting the algorithms to optimality criteria for model selection, possibly combined with D- and I-criteria.

Finally, another challenging future development could be the extension of the proposed algorithms to the generalised linear model, because the definition of outliers and high leverage points, in this context, is not straightforward.

Acknowledgements We are grateful to Prof. Claudio Agostinelli of the University of Trento, for the useful discussions related to robust statistics, which has stimulated the development of this study.

Funding Open access funding provided by Università Cattolica del Sacro Cuore within the CRUI-CARE Agreement.

Appendix

The following algorithm provides a "good" initial sample for Algorithms 2 and 4.

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 $j = 1, \ldots, \tilde{N}$, 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

References

Atkinson A, Donev A, Tobias R (2007) Optimum experimental designs, with SAS. Oxford University Press, Oxford

Chatterjee S, Hadi AS (1986) Influential observations, high leverage points, and outliers in linear regression. Stat Sci 1(3):379–416

Deldossi L, Tommasi C (2022) Optimal design subsampling from Big Datasets. J Qual Technol 54(1):93–101

Drovandi CC, Holmes CC, McGree JM, Mengersen K, Richardson S, Ryan EG (2017) Principles of experimental design for big data analysis. Stat Sci 32(3):385–404

Fedorov VV (1972) Theory of optimal experiments. Academic Press, New York

Harman R, Filová L (2019) OptimalDesign: a toolbox for computing efficient designs of experiments. R package version 1.0.1. https://CRAN.R-project.org/package=OptimalDesign

Hoaglin DC, Welsch RE (1978) The hat matrix in regression and ANOVA. Am Stat 32(1):17–22

Ma P, Mahoney MW, Yu B (2015) A statistical perspective on algorithmic leveraging. J Mach Learn Res 16(27):861–911

Meyer RK, Nachtsheim CJ (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. Technometrics 37(1):60–69

Searle SR (1982) Matrix algebra useful for statistics. Wiley, New York

Wang H, Yang M, Stufken J (2019) Information-based optimal subdata selection for Big Data linear regression. J Am Stat Assoc 114(525):393–405