Computational outlier detection methods in sliced inverse regression

Hadrien Lorenzo, Jérôme Saracco

Computational outlier detection methods in sliced inverse regression

Hadrien Lorenzo and Jérôme Saracco

Abstract Sliced inverse regression (SIR) focuses on the relationship between a dependent variable y and a p-dimensional explanatory variable x in a semiparametric regression model in which the link relies on an index $x'\beta$ and link function f. SIR allows to estimate the direction of β that forms the effective dimension reduction (EDR) space. Based on the estimated index, the link function f can then be nonparametrically estimated using kernel estimator. This two-step approach is sensitive to the presence of outliers in the data. The aim of this paper is to propose computational methods to detect outliers in that kind of single-index regression model. Three outlier detection methods are proposed and their numerical behaviors are illustrated on a simulated sample. To discriminate outliers from "normal" observations, they use IB (in-bags) or OOB (out-of-bags) prediction errors from subsampling or resampling approaches. These methods, implemented in R, are compared with each other in a simulation study. An application on a real data is also provided.

1 Introduction

On one hand, classical linear regression or more generally parametric regression have achieved resounding success in many real problems whose goal is to investigate the relationship between a response variable $y \in \mathbb{R}$ and a covariate $x \in \mathbb{R}^p$. However it can be argued that assuming specific structural constraints on the link function of y on x is too stringent. On the other hand, nonparametric regression is clearly a more flexible approach, but it is well-known that it typically suffers from the curse of dimensionality, i.e., a poor rate of convergence when the dimension p of

Hadrien Lorenzo
Inria BSO, 33400 Talence, France, e-mail: hadrien.lorenzo@inria.fr

Jérôme Saracco
Inria BSO & ENSC Bordeaux INP, 33400 Talence, France, e-mail: jerome.saracco@ensc.fr

x increases. To address these problems from purely parametric or nonparametric approaches, several authors studied single-index or multiple-index models. This kind of regression models can be viewed as an alternative semiparametric approach based on sufficient dimension reduction. So, in a dimension reduction setting, many authors suppose that x can be replaced by a linear combination of its components, $\beta' x$, without losing information on the conditional distribution of y given x. One way to express this assumption is:

$$y \perp x \mid \beta' x \tag{1}$$

where the notation $v_1 \perp v_2 \mid v_3$ means that the random variable v_1 is independent of the random variable v_2 given any values for the random variable v_3. One can write (1) as, for instance, the following single-index model with an additive error:

$$y = f(\beta' x) + \varepsilon, \tag{2}$$

where f is an unknown real-valued function, the distribution of ε is arbitrary and unknown, and $\varepsilon \perp x$. Since f is unknown, the p-dimensional parameter β is not totally identifiable, but the subspace spanned by β is identifiable. This subspace is referred to as the effective dimension reduction (EDR) subspace following Duan and Li [15] in their original presentation of sliced inverse regression (SIR). Li [26] consider a multiple-index regression model. The Euclidean parameter β is now a $p \times K$ matrix: $\beta = [\beta_1, \ldots, \beta_K]$ where the vectors β_k are assumed to be linearly independent. The EDR subspace is then the K-dimensional linear subspace of $\mathbb{R}^p$ spanned by the β_k's.

Note that the dimension reduction is very useful in an exploratory stage of data analysis since model (1) relies on very few structural assumptions. For instance, it is not assumed that the indices act additively as often assumed in multiple-index models. It is likewise not necessary to assume that the error term is additive on mean (as for the model (2)), thus heteroscedastic model are potentially included in this modelling. Note also that sufficient dimension reduction of the regression leads to summary plot of y versus estimated indices which provides useful graphical modelling information.

In a second step, to study the relationship between the response variable and the few estimated indices, standard nonparametric approaches (such as kernel or spline smoothing) can be used. This stage usually involves additional assumptions such as an additive error term (as in model (2)) to get consistent properties of the corresponding estimate of the link function f.

In the statistical literature, different methods have been developed with the aim of estimating the EDR subspace. SIR, SIR-II, SIR$_\alpha$, SAVE (sliced average variance estimation) and pHd (principal Hessian directions) approaches are the most popular, see [2, 5, 6, 7, 9, 18, 21, 22, 25, 27, 28, 29, 34, 35, 38, 39, 40] among others. The important question of the determination of the EDR space dimension in SIR and related methods has also been much studied, see for example [17, 30].

SIR is known to be a relevant technique for the purpose of dimension reduction. Several properties of SIR have been extensively studied and numerous extensions have been already proposed. However little attention has been paid to sensitivity of SIR to outliers or to robustness aspects. Since SIR theory is based on conditional expectation and covariance matrix properties (see Sect. 2 for details), it is obvious that SIR can be severely influenced by outliers in the data, see [19] or [10] for instance. In [32, 33], the detection of influential observations on the estimation of the dimension reduction subspaces returned by SIR, SIR-II and SAVE have been studied using the notion of influence functions of single observations. However, the proposed empirical influence values are very sensitive to the choice of the number H of slices (introduced in the next section) in detecting influential observations, which makes this approach complicated to use in practice. Robust SIR methods were then developed and only focused on the estimation of the EDR space (regardless of the estimation of the link function f). For example, in [8] the inverse regression formulation of SIR is therefore extended to non-Gaussian errors with heavy-tailed distributions (Student). The underlying Expectation-Maximization algorithm was tested in presence of outliers and provided good numerical results. [14] also mentioned that classical sufficient dimension reduction methods are sensitive to outliers present in predictors, and may not perform well when the distribution of the predictors is heavy-tailed. Two robust inverse regression methods which are insensitive to data contamination (weighted inverse regression estimation and sliced inverse median estimation) were then introduced and they demonstrated very interesting numerical performances in the presence of potential outliers. In the same spirit, [3] proposed sliced inverse median difference regression to robustify SIR methodology at the presence of outliers. In [13], robust SIR extensions were presented through robust estimates of the covariance matrix.

The goal of this paper is to propose computational methods to detect outliers in a single-index regression model, comprising EDR space estimation using SIR and link function estimation based on kernel smoothing. In practice, it is always interesting to detect outliers (rather than only developing robust methods), to isolate them and to understand why these observations are aberrant (wrong numerical values, unusual individuals, ...). Once the dataset has been cleaned, it is then possible to implement the usual methodology, SIR followed by a non-parametric estimation of f.

In Sect. 2 a brief overview of usual SIR is given. Three outlier detection methods, named MONO, TTR and BOOT hereafter, are presented in Sect. 3. They use IB (in-bags) or OOB (out-of-bags) prediction errors from subsampling or resampling approaches in order to discriminate outliers from "normal" observations. These methods have been implemented in R. How these methodologies work is described on a simulated example in Sect. 4. Sect. 5 provides a more extensive simulation study that compares the numerical performances of the proposed methods. A real dataset is also used to illustrate these approaches in Sect. 6. Finally concluding remarks are given in Sect. 7.

2 A brief review of usual SIR

In order to estimate the EDR space, various methods based on the use of inverse regression are widely available in literature. In order for inverse regression to be useful in estimating the EDR space, some of them, like SIR or SAVE or principal Hessian direction, need additional conditions on the marginal distribution of the covariate x. In this paper, we focus the usual SIR approach which relies on the following linearity condition (LC) on x:

$$\text{For all } b \in \mathbb{R}^p, \mathbb{E}\left[b'x \mid \beta'x\right] \text{ is linear in } x'\beta. \tag{3}$$

Note that the LC is required to hold only for the true Euclidean parameter β. Since β is unknown, it is not possible in practice to verify a priori this assumption. Therefore we can assume that LC holds for all possible values of β, this is equivalent to assume an elliptical symmetry of the distribution of x: for instance the well-known multivariate normal distribution satisfies this condition. Finally, following [20], the LC is not a severe restriction because this LC holds to a good approximation in many problems as the dimension p of the predictors increases. Interesting discussions on the LC can also be found in [7, 25] for instance.

Let us now consider a monotone transformation T. Under model (1) and LC, [15] showed that the centered inverse regression curve satisfies:

$$\mathbb{E}[x \mid T(y)] - \mu \in \text{Span}(\Sigma\beta), \tag{4}$$

where $\mu := \mathbb{E}[x]$ and $\Sigma := \mathbb{V}(x)$. Therefore the space spanned by the centered inverse curve, $\{\mathbb{E}[x \mid T(y)] - \mathbb{E}[x] \ : \ y \in \mathcal{Y}\}$ where $\mathcal{Y}$ is the support of response variable y, is a subspace of the EDR space, but it does not guarantee equality. A pathological model, often called symmetric dependent model, has been identified in the literature, and is model for which the centered inverse regression curve is degenerated. To solve this problem, specific methods (based on higher order inverse moments), such as SIR-II, SIR$_\alpha$ or SAVE, have been developed.

When the model is not pathological (which is often the case in pratice), the centered inverse regression curve can be used to recover the EDR space from (4). Indeed, a direct consequence of this result is that the covariance matrix of this curve,

$$\Gamma := \mathbb{V}(\mathbb{E}[x \mid T(y)]),$$

is degenerate in any direction Σ-orthogonal to β (i.e. to the β_k's for a multiple-index model). Therefore, the eigenvectors associated with the non null eigenvalues of $\Sigma^{-1}\Gamma$ are EDR directions, which means that they span the EDR space E.

In the slicing step of SIR, the range of y is partitioned into H non-overlapping slices $\{s_1, \ldots, s_H\}$. With such slicing, the covariance matrix Γ can be straightforwardly written as

$$\Gamma := \sum_{h=1}^{H} p_h (m_h - \mu)(m_h - \mu)'$$

where $p_h = P(y \in s_h)$ and $m_h = \mathbb{E}[x \mid y \in s_h]$.

Let us now consider a random sample $\{(x_i, y_i), i = 1, \ldots, n\}$ generated from the single-index regression model (2). By substituting the empirical versions of μ, Σ, p_h and m_h for their theoretical counterparts, we obtain an estimated basis of E spanned by the eigenvector $\hat{b}_{\text{SIR}}$ associated with the largest eigenvalue of the estimate $\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n$ of $\Sigma^{-1}\Gamma$ where

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)' \text{ and } \widehat{\Gamma}_n = \sum_{h=1}^{H}\hat{p}_{h,n}(\hat{m}_{h,n} - \bar{x}_n)(\hat{m}_{h,n} - \bar{x}_n)',$$

with $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\hat{n}_{h,n} = \sum_{i=1}^{n}\mathbb{I}_{[y_i \in s_h]}$, $\hat{p}_{h,n} = \frac{\hat{n}_{h,n}}{n}$, $\hat{m}_{h,n} = \frac{1}{\hat{n}_{h,n}}\sum_{i \in s_h} x_i$, the notation $\mathbb{I}_{[.]}$ standing for indicator function. This approach is the one proposed by [15, 26] when they initially introduced the SIR approach. Since the early 1990s, the SIR method has been extensively studied by many authors, see for instance all the references mentioned in the introduction.

The link function f of model (2) can then be estimated by the usual kernel estimator (see for example [36]) based on the sample $\{(x_i'\hat{b}_{\text{SIR}}, y_i),\ i = 1, \ldots, n\}$ where the $x_i'\hat{b}_{\text{SIR}}$'s are the values of the estimated index. For a given value x_0 of x, the kernel estimation of $f(\beta' x_0)$ is given by

$$\hat{f}_n(\hat{b}_{\text{SIR}}' x_0) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i'\hat{b}_{\text{SIR}} - x_0'\hat{b}_{\text{SIR}}}{h_n}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i'\hat{b}_{\text{SIR}} - x_0'\hat{b}_{\text{SIR}}}{h_n}\right)},$$

where K is the kernel and h_n is the bandwidth. The kernel is usually a positive symmetric weighting function with integral equal to 1. In the rest of the paper, the chosen kernel is the density of the normal distribution $\mathcal{N}(0, 1)$, called the Gaussian kernel. The bandwidth $h_n > 0$ is called the smoothing parameter in kernel regression because it controls variance and bias of the estimator. This parameter must therefore be correctly tuned using cross-validation for instance:

$$h_n^{\text{opt}} = \arg\min_{h_n > 0} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_n^{(-i)}(\hat{b}_{\text{SIR}}' x_0)\right)^2,$$

where $\hat{f}_n^{(-i)}(\hat{b}_{\text{SIR}}' x_0)$ stands for the estimation of $f(\beta' x_0)$ based on the sample $\{(x_j'\hat{b}_{\text{SIR}}, y_j),\ j \neq i\}$.

When the underlying regression model is a multiple-index model, the estimated EDR space is spanned by the eigenvectors associated with the largest K eigenvalues of the estimate $\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n$. Let $\widehat{B}_{\text{SIR}}$ be the $p \times K$ matrix of these K eigenvectors. The estimated indices $x_i'\hat{B}_{\text{SIR}}$'s are now K-dimensional and the kernel estimation of $f(\beta' x_0)$ is then based on multivariate kernel. For example K can be the density of the multivariate normal distribution $\mathcal{N}(0_K, I_K)$ where 0_K (resp. I_K) stands for the null vector of dimension K (resp. the identity matrix of order K), and the associated smoothing

parameter h_n can be K-dimensional. Another way is to consider the following kernel estimator

$$\hat{f}_n(\hat{B}'_{\text{SIR}}x_0) = \frac{\sum_{i=1}^{n} K\left(\frac{||x'_i\hat{B}_{\text{SIR}} - x'_0\hat{B}_{\text{SIR}}||}{h_n}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{||x'_i\hat{B}_{\text{SIR}} - x'_0\hat{B}_{\text{SIR}}||}{h_n}\right)},$$

where $||.||$ stands for a chosen norm in $\mathbb{R}^K$ and the bandwidth h_n is unidimensional.

3 Outlier detection methods in SIR

Three outlier detection methods for single-index regression model (2) are presented. Let us consider a sample $S = \{(x_i, y_i),\ i = 1, \ldots, n\}$ of n individuals among which some may be outliers.

For each of the three methods, the parameter β (more properly, the EDR direction b) is estimated by the usual SIR method (with the number of slices $H = 10$) and the link function f is estimated using the kernel estimator with the Gaussian kernel and the bandwidth tuned via cross-validation.

3.1 A naive method

This naive method relies on the following three steps.

Step 1. Estimation the EDR direction from the sample S.

The usual SIR provides the estimate $\hat{b}_{\text{SIR}}$ of b. The corresponding indices $\{\hat{b}'_{\text{SIR}}x_i,\ i = 1, \ldots, n\}$ are then calculated.

Step 2. Estimation of the adjusted value $f(\beta' x_i)$'s.

From the sample $\{(\hat{b}'_{\text{SIR}}x_i, y_i),\ i = 1, \ldots, n$, the adjusted values are obtained via the kernel estimator based on the Gaussian kernel and the bandwidth tuned via cross-validation. Let $\hat{y}_i = \hat{f}_n(\hat{b}'_{\text{SIR}}x_i)$ for $i = 1, \ldots, n$.

Step 3. Evaluation of the error associated with the model estimation and outlier detection.

The errors considered are naturally the residuals: for $i = 1, \ldots, n$, $\hat{e}_i = y_i - \hat{y}_i$.
The detection of potential outliers is simply based on the definition of outliers in the boxplot of the absolute error $|\hat{e}_i|$'s, i.e. the outliers correspond to individuals whose values are greater than the value of the 3rd quartile plus 1.5 times the inter-quartile interval.
Note that, in the same spirit, the bootstrap histogram of "mean - trimmed mean" for a suitable trimming number was proposed by [37] as a nonparametric graphical tool for detecting outlier(s) in a dataset. The bootlier-plot was introduced and it is shown that the multimodality in the bootlier plot is caused by outlier(s) in the sample.

This naive method is called MONO hereafter. The name MONO stands for a single use of the initial sample S and a single estimate of the underlying single-index model. In the numerical example of Sect. 4, Fig. 1 allows to visualize the position of the outliers in the corresponding boxplot.

3.2 TTR method

This method relies on training sample and test sample replications for evaluating the "stability" of the estimated model, hence the name TTR of the method for Training Test Replications.

The TTR approach works in two major steps. Let R be the number of replications chosen by the user. In practice $R = 2000$ is more than enough for reasonable sample sizes, i.e. $n \leq 500$. Let $\alpha \in [0, 1]$ be the proportion of the sample which will constitute the test sample. In the rest of the paper, the parameter is fixed to $\alpha = 0.1$, thus 90% of the sample S is used as the training sample S_{train} and the remaining 10% of the sample S constitutes the test sample S_{test}. Note that individuals are drawn with equal weight and without replacement.

Step 1. For each replication r (with $r = 1, \ldots, R$),

1.a. Split the initial sample S into a training sample $S_{\text{train}}^{(r)}$ and a test sample $S_{\text{test}}^{(r)}$ containing respectively $(1 - \alpha)$% and α% of the individuals.

1.b. Using $S_{\text{train}}^{(r)}$, calculate the estimated EDR direction $\hat{b}_{\text{SIR}}^{(r)}$ and the associated indices $\{(\hat{b}_{\text{SIR}}^{(r)})' x_i,\ i \in S_{\text{train}}^{(r)}\}$.

1.c. For all the individuals $i^* \in S_{\text{test}}^{(r)}$, calculate the error of prediction of the response variable y as follows:

$$\hat{e}_{i^*}^{(r)} = y_{i^*} - \hat{f}_n^{(r)} \left((\hat{b}_{\text{SIR}}^{(r)})' x_{i^*} \right),$$

where the estimate $\hat{f}_n^{(r)}(.)$ is based on the sample $\{((\hat{b}_{\text{SIR}}^{(r)})' x_i, y_i),\ i \in S_{\text{train}}^{(r)}\}$.

Step 2. Evaluation of the error means.

For each $i^* = 1, \ldots, n$, calculate the associated error mean over the R replications (when the individual i^* is present in the corresponding test sample):

$$\overline{e}_{i^*} = \frac{\sum_{r=1}^{R} \left| \hat{e}_{i^*}^{(r)} \right| \mathbb{I}_{[i^* \in S_{\text{test}}^{(r)}]}}{\sum_{r=1}^{R} \mathbb{I}_{[i^* \in S_{\text{test}}^{(r)}]}}.$$

Step 3. Detection of the outliers via a change point detection.

The idea is to find a single change point position in the sequence of the errors' means $\{\overline{e}_{(i^*)},\ i^* = 1, \ldots, n\}$ ordered by decreasing values (where the subscript (i^*) enclosed in parentheses indicates the i^*th order statistic of the sample). Indeed,

if there are no outlier in the data, no change points should clearly appear in this sequence of ordered absolute mean errors. On the other hand, in the presence of outliers, the corresponding mean absolute errors should naturally be significantly larger than the errors associated with other individuals. Thus, looking for a single change point in mean and variance in this sequence should intuitively allow us to separate outliers from other observations.

Many authors have proposed a single search method to detect change points. Recently, [23] have developed the R package `changepoint` that helps to detect the location of different change points. For single or multiple change point detection, the approach allows to estimate the points at which the statistical properties of a sequence of observations change. Within this package, several change in mean methods are available as well as methods focusing on detection of change in variance and methods searching a change in both mean and variance. Briefly, let us give an overview of the underlying approach. Let $z_{1:n} = (z_1, \ldots, z_n)$ be the ordered sequence of the errors' means and $\tau_{i:m} = (\tau_1, \ldots, \tau_m)$ the positions of the m change points (each change point position is between 1 and $n-1$, $\tau_0 = 0$ and $\tau_{m+1} = n$). The idea is to minimize

$$\sum_{i=1}^{m+1} [C(z_{(\tau_{i-1}+1):\tau_i})] + \gamma g(m) \tag{5}$$

where C is a cost function (for instance negative log-likelihood ratio test statistic) and $\gamma g(m)$ is a penalty to guard against over fitting. This package implements several algorithms to minimize (5): binary segmentation [16], segment neighborhood [1] and pruned exact linear time (PELT) [24]. Here the `changepoint` package is used to detect only one change point ($m = 1$) in mean and variance with the binary segmentation algorithm in the ordered sequence of means $\{\overline{e}_{(i^*)},\ i^* = 1, \ldots, n\}$. In the numerical example of Sect. 4, Fig. 2 (top left) visualizes the position of the estimated single change point.

An individual associated with an ordered error's mean before the single change point position is then considered as an outlier.

Remark. In the associated R code, the bandwidth is tuned only once in step 1.c for the kernel estimation of each iteration. This "optimal" bandwidth is obtained via cross-validation using the whole sample of the y_i's versus the estimated indices $x_i'\hat{b}_{\text{SIR}}$. This is a reasonable choice if one assumes that there are no outlier in the x_i's and thus in the $x_i'\hat{b}$'s or in the $x_i'\hat{b}^{(r)}$'s. Note that the presence of visible outliers in the x_i's would have been detected in a preliminary step and the dataset would then have been cleaned. This choice of only one tuned bandwidth clearly saves calculation time for the TTR method. Finally, note also that, in each iteration of step 3 for the TTR method, it is easy to integrate an automatic optimal bandwidth selection in the R code. The same strategy is used for the BOOT method in step 1.c presented in the following section.

3.3 BOOT method

The MONO method deals with in-bag (IB) errors and the TTR method with out-of-bag (OOB) errors. While MONO risks overfitting, TTR risks significance loss of statistical power (since the train sample is a subsample of the initial sample) but cannot quantify the impact of IB individuals. The current BOOT method uses IB errors in that objective.

Isolated individuals that are not outliers, in the plot of the estimated index versus the response y are usually hard to predict especially if those individuals are not in the training dataset. However, if any of those individuals are included in the training dataset, it has a beneficial effect on the built model. Indeed, those individuals are therefore better predicted while the non isolated individuals are still well predicted since those isolated individuals are in line with the regression model. For those isolated individuals, the OOB error is then high while the IB error is potentially low. They are denoted as "borderline" observations in the following. On the other hand, the "outliers" are always badly predicted with high IB and OOB errors and the "normal" individuals are always well predicted with low IB and OOB errors (see an illustration of these comments in Figure 4 that gives examples of those three types of observations).

The BOOT method is based on two simple decision rules to discriminate between these three types of individuals ("normal" observation, "borderline" observation,"outlier") using the IB error and its logarithmic transformation. This method relies on bootstrap samples of S. Let B be the number of bootstraps chosen by the user. In practice $B = 2000$ is more than enough for reasonable sample sizes, i.e. $n \leq 500$. Note that individuals are drawn with equal weight and with replacement.

Step 1. For $b = 1, \ldots, B$,

1.a. Draw a bootstrap sample $S^{(b)}$ from the initial sample S. Let $n_i^{(b)}$ denote the number of times the observation i is present in the bootstrap sample $S^{(b)}$.

1.b. Using $S^{(b)}$, calculate the corresponding estimated EDR direction $\hat{b}_{\text{SIR}}^{(b)}$ and the associated indices $\{(\hat{b}_{\text{SIR}}^{(b)})' x_i,\ i \in S^{(b)}\}$.

1.c. For all the individuals $i \in S^{(b)}$, calculate the IB error of prediction of the response variable y as follows:

$$\hat{e}_i^{(b)} = y_i - \hat{f}_n^{(b)} \left((\hat{b}_{\text{SIR}}^{(b)})' x_i \right),$$

where the estimate $\hat{f}_n^{(b)}(.)$ is based on the sample $\{((\hat{b}_{\text{SIR}}^{(b)})' x_i, y_i),\ i \in S^{(b)}\}$. Note that, even if they are not used in Steps 2 and 3, the OBB errors (for all the individuals $i \notin S^{(b)}$) have also been calculated since they are used in graphical representations (see Fig.4).

Step 2. Evaluation of the error means.

For each $i = 1, \ldots, n$, calculate the associated error mean over the B replications (when the individual i is present at least once in the corresponding bootstrap

sample):

$$\overline{\overline{e}}_{(i)} = \frac{\sum_{b=1}^{B} \left| \hat{e}_i^{(b)} \right| \mathbb{I}_{[i \text{ such that } n_i^{(b)} \geq 1]}}{\sum_{b=1}^{B} \mathbb{I}_{[i \text{ such that } n_i^{(b)} \geq 1]}}.$$

Step 3. Detection of outliers and "bordeline" observations

The idea here is to first identify among the errors $\{\overline{\overline{e}}_{(i)},\ i = 1, \ldots, n\}$ those which are particularly high and which will naturally correspond to these "big" outliers. For this purpose, the log scale was used to detect these outliers. Then, in a second step, the usual scale is used in order to identify other possible "small" remaining outliers which are then called "borderline" observations.

3.a. The detection of potential outliers is based on the definition of outliers in the boxplot[1] of the $\log\left(\overline{\overline{e}}_{(i)}\right)$'s. The corresponding observations are plotted in blue in Fig. 3 (on the left).

3.b. The detection of potential "borderline" observations is based on the definition of outliers in the boxplot of the $\overline{\overline{e}}_{(i)}$'s, these "current outliers" are plotted with orange triangle in Fig. 3 (in the middle). The "borderline" observations are thus defined as the current detected outliers not identified as outliers in the previous step 3.a. (plotted with blue circle behind the orange triangle in this graphic). The corresponding "borderline" observations are therefore those represented only in orange on the graphic in Fig. 3 (on the right).

Remark. In Step 3.a., the log transformation is used by default to detect the potential outliers. However, the relevant transformation of the considered errors, $\overline{\overline{e}}_{(i)},\ i = 1, \ldots, n$, is probably not always log but that it may depend on the link function f itself and on the distribution of ϵ in the regression model (2).

4 A numerical example

Let us consider a simulated sample to clearly illustrate how the previous three outlier detection methods (MONO, TTR and BOOT) work. Note that steps 3 of the different methods (MONO, TTR and BOOT) are interchangeable with each other and thus they can be used after any of the error calculation steps (steps 1 and 2). In Sections 4 and 5, only the MONO, TTR and BOOT methods are compared with each other, not ideally all possible combinations. Thus, we are well aware that this will make it difficult to identify whether the success of the method is due mainly to the different error calculation processes (steps 1 and 2) or to the technique of detecting "abnormally large" errors (step 3).

[1] already described in the presentation of the MONO method

4.1 Description of the simulated dataset

The following single-index regression model is used in all numerical studies in Sections 4 and 5:

$$y = \frac{(x'\beta)^3}{100} + \epsilon, \tag{6}$$

where $\beta = (2, 2, 1, -2, -3, 0, \ldots, 0)' \in \mathbb{R}^p$, x follows the p-dimensional uniform distribution on $[-2; 2]^p$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 0.25)$ is independent of x. In a first step, $\tilde{n} = 200$ observations $\{(x_i, y_i),\ i = 1, \ldots, \tilde{n}\}$ are generated from model 6 with $p = 5$. Then in a second step, $\tilde{\tilde{n}} = 10$ new individuals are generated as follows: for $i = \tilde{n} + 1, \ldots, \tilde{n} + \tilde{\tilde{n}}$,

- x_i is drawn from the uniform distribution on $[-2; 2]^p$,
- y_i is drawn (independently from x_i) from the uniform distribution on the support of the first $\tilde{n}$ values of y.

These $\tilde{\tilde{n}}$ new observations are then "potential" outliers for the model (6) since their y_i's are not linked to the x_i's via this model. Note that these observations are not outliers regarding the distribution of the x_i's (resp. of the y_i's). The term "potential" refers to the fact that an observation (x_i, y_i) (for an $i \in \{\tilde{n} + 1, \ldots, \tilde{n} + \tilde{\tilde{n}}\}$ may be close, just by chance, to the "true" structure of the data (based on the underlying model (6)). The objective is to detect these potential $\tilde{\tilde{n}}$ outliers in the sample $S = \{(x_i, y_i),\ i = 1, \ldots, n\}$ where $n = \tilde{n} + \tilde{\tilde{n}}$ and then to estimate as best as possible the relationship between y and x through the single-index $x'\beta$.

4.2 Numerical results

In a first step, based on the available sample $S = \{(x_i, y_i),\ i = 1, \ldots, n\}$, the EDR direction b is estimated by $\hat{b}_{\text{SIR}}$ using the usual SIR method (with the number of slices $H = 10$) and the link function f is estimated by $\hat{f}_n(.)$ using the kernel estimator with the Gaussian kernel and the bandwidth tuned via cross-validation. The distance between the true EDR space and the estimated one is defined as

$$d^2(E, \widehat{E}) = 1 - \frac{\text{Trace}(P_E P_{\widehat{E}})}{K} \in [0, 1],$$

where $P_E = \beta(\beta'\beta)^{-1}\beta'$ (resp. $P_{\widehat{E}}$) is the orthogonal projector onto E (resp. $\widehat{E}$) with K the dimension of the EDR space (here $K = 1$ for a single-index model). The closer this distance is to zero, the better the estimation $\widehat{E}$ of E. On the simulated sample, we have $d^2(E, \widehat{E}) = 0.0093$. The corresponding MSE (mean squared error) defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_n(x_i'\hat{b}_{\text{SIR}})\right)^2$$

is equal to $MSE = 12.99$.

Using the naive MONO method, 15 outliers have been detected, see Fig. 1 (left) for the boxplot of the absolute residual errors (with outliers in blue) and Fig. 1 (right) for the visualization of these outliers on the plot of the estimated indices $x_i'\hat{b}_{\text{SIR}}$ versus the y_i's. Note that all the $\tilde{\tilde{n}} = 10$ generated outliers have been identified. Among these 15 detected outliers, 5 are false positive, however the individual 21 (at the top right of the plot of Fig. 1(right)) can be considered as an "extreme" observation. An "extreme" observation may obviously be detected as an outlier by the method because the nonparametric estimation of f by the kernel method is based on local smoothing. Thus since an "extreme" observation is too isolated in the plot of the estimated indices $(x_i'\hat{b}_{\text{SIR}},\ i = 1, \ldots, n)$ versus the y_i's, its kernel prediction is difficult due to the lack of observations around it (this is the problem of data sparsity in nonparametric regression). Using the initial sample without these 15 outliers, the associated MSE is now equal to 0.24, and we have $d^2(E, \widehat{E}) = 0.00361$. These two quantities clearly show the benefits of removing the detected outliers.

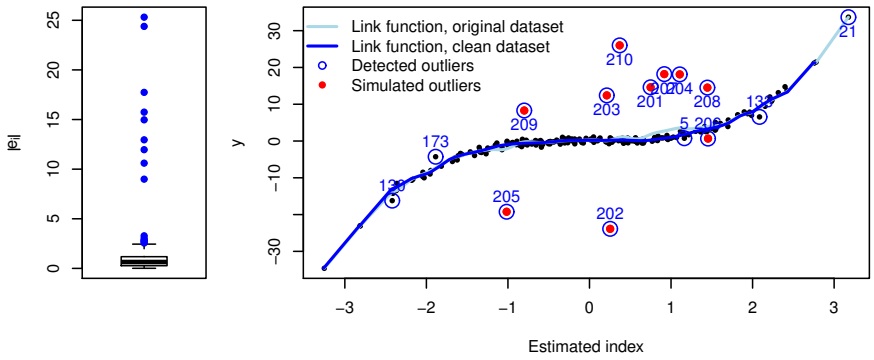

Fig. 1 MONO method description. Left graphic provides the boxplot of the absolute errors, the detected outliers are in blue. On the plot of the estimated indices $x_i'\hat{b}_{\text{SIR}}$ versus the y_i's (right graphic), these outliers are also plotted in blue, red points correspond to the $\tilde{\tilde{n}}$ (true) potential outliers. The kernel estimations of the link function are superimposed for both the original dataset (in light blue) and the dataset without the detected outliers (in dark blue).

Using the TTR method with $R = 3000$, 11 outliers have been detected, see Fig. 2 (top left) for the detection of the unique change point position in the sequence of the ordered errors' means, $\{\overline{e}_{(i^*)},\ i^* = 1, \ldots, n\}$. Fig. 2 (top right) provides the visualization of these outliers on the plot of the estimated indices $x_i'\hat{b}_{\text{SIR}}$ versus the y_i's. Among these 11 outliers, only 2 are false positive: observations 10 and 21 (at the bottom left and at the top right of the plot of Fig. 2 (top right)) can naturally be considered as "extreme" observations but they are still selected as outliers for the same reasons of nonparametric kernel estimation as those mentioned for the MONO method. Note also that observation 206 (in red) has not been detected as an outlier by TTR method, but its projection is very close to the "true data" (in black, i.e. that is those generated by the underlying model) and thus this observation is not

really a significant outlier. Using the initial sample without these 11 outliers, the associated MSE is now equal to 0.29, and we have $d^2(E, \widehat{E}) = 0.00367$. The benefits of removing these detected outliers is again very clear. Fig. 2 (bottom) provides the plot of the estimated indices $x_i'\hat{b}_{\text{SIR}}$ versus the y_i's considering the dataset without the detected outliers. The kernel estimation of the link function (in blue) is superimposed on the plot. One can observe the very good fit of the data to the underlying model.

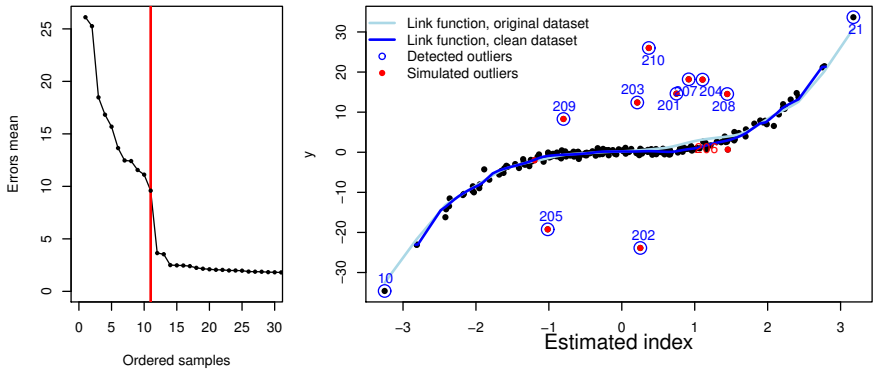

Fig. 2 TTR method description. Top-left graphic shows the ordered means errors with the red vertical line providing the estimated single change point position. On the plot of the estimated indices $x_i'\hat{b}_{\text{SIR}}$ versus the y_i's (top-right graphic), these outliers are also plotted in blue, red points correspond to the $\tilde{\tilde{n}}$ (true) potential outliers. The kernel estimations of the link function are superimposed for both the original dataset (in light blue) and the dataset without the detected outliers (in dark blue).

Using the BOOT method with $B = 3000$, 9 out of the $\tilde{\tilde{n}} = 10$ outliers were detected, and 4 "bordeline" observations have been identified. Fig. 3 (right) provides the visualization of the outliers (in blue) and of the "borderline" observations (in orange) on the plot of the y_i's versus the estimated indices $x_i'\hat{b}_{\text{SIR}}$. The boxplot on the right allows to detect the outliers while the boxplot in the middle identifies the "borderline" observations. The individual 206 (simulated as an outlier) is here detected as a "bordeline" observation. Note that there is no false positive. Graphics in Fig. 4 provide the plot of the $n_i^{(b)}$'s versus the $|e_i^{(b)}|$'s (for $b = 1, \ldots, B$) for three individuals. The horizontal line on each plot represents the corresponding error mean $\overline{|e_{(i)}|}$ over the B replications (when the individual i was present at least once in the corresponding bootstrap sample). One can observe that:

- for observation 1 (which is a "normal" observation), the corresponding mean $\overline{|e_{(1)}|}$ is low,
- for observation 21 (which is characterized as a "borderline" observation"), the corresponding mean $\overline{|e_{(21)}|}$ is intermediate. The model learns its position and modifies its tail, which explains the fall in error between $n_{(21)}^{(b)} = 0$ and $n_{(21)}^{(b)} = 1$,

- for observation 209 (which was detected as an outlier), the corresponding mean $\overline{\overline{|e_{(209)}|}}$ is clearly higher than the previous ones, no matter the number of times that observation is present in the bootstrap sample.

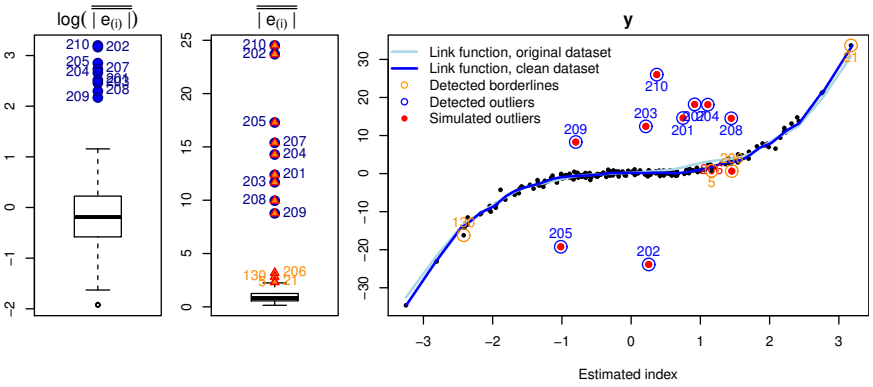

Fig. 3 BOOT method description. The two graphics on the left correspond to the boxplots of the *log* mean absolute errors (defining outliers, in blue) and of the mean absolute errors (defining "borderline" observations, in orange). Selected outliers (in blue) and selected "borderline" observations (in orange) are showed on the plot of the estimated indices $x_i'\hat{b}_{\text{SIR}}$ versus the y_i's (right graphic), the red points correspond to the $\tilde{\tilde{n}}$ (true) potential outliers. The kernel estimations of the link function are superimposed for both the original dataset (in light blue) and the dataset without the detected outliers and " borderline" observations in dark blue).

Using the initial sample without these 9 outliers and 4 "bordeline" observations, the associated MSE is now equal to 0.32 and we have $d^2(E, \widehat{E}) = 0.00172$, which highlights the high effectiveness of the BOOT method. Finally, let us remark that, in the computation of the mean descriptors, we chose to consider only the (absolute) error values for which each individual is represented at least once in the bootstrap sample as to prevail from selecting "extreme" observations, as discussed in the comments of the previous two methods.

5 Simulation results

In this simulation study, $N = 100$ replications of samples from model (6) have been generated with various values of the sample size $\tilde{n}$ (= 100, 200, 300), various values of the dimension p (= 5, 20) of the covariate x, and two numbers of potential outliers $\tilde{\tilde{n}}$ (= 3, 10). For each generated sample and each outlier detection method (MONO, TTR with $R = 2000$ and BOOT with $B = 2000$), the following quantities were calculated:

- the quality of the estimated EDR direction $d^2(E, \widehat{E})$ where $\widehat{E}$ is the estimated EDR space based on the complete sample (unique for all the three methods),

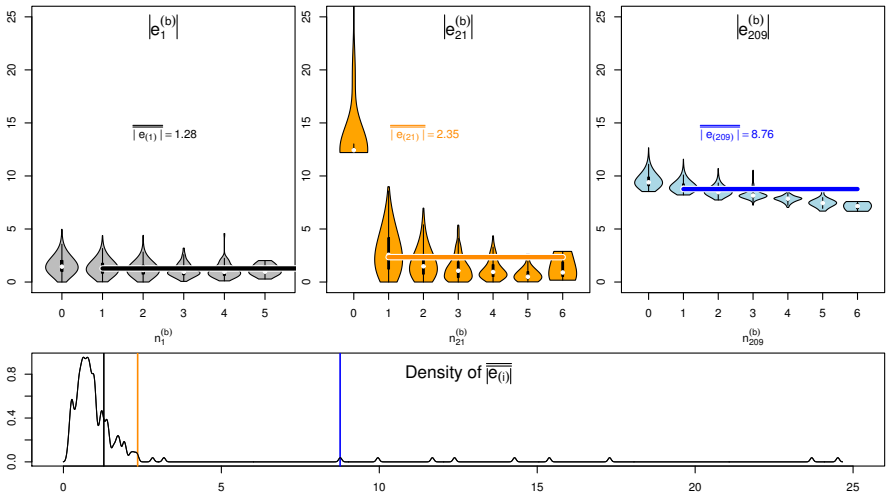

Fig. 4 For the BOOT method, plots of the $|e_{(i)}^{(b)}|$'s versus the $n_i^{(b)}$'s (for $b = 1, \ldots, B$) for three individuals: a "normal" individual ($i = 1$) on the left, a "borderline" individual ($i = 21$) in the middle, and an outlier ($i = 209$) on the right. Colored (resp. black, orange and blue) segments show their corresponding computed IB error means $\overline{\overline{|e_{(i)}|}}$ (for $n_i^{(b)} \geq 1$). The last plot at the bottom provides a density estimation of the $\overline{\overline{|e_{(i)}|}}$'s with these three individuals showed through colored vertical lines. Since the OOB errors (for $n_i^{(b)} = 0$) are not used, individual $i = 21$ (in orange) is not considered as outlier but as "borderline" observation.

- the MSE evaluated on the complete sample (unique for all the three methods),
- the number of detected outliers (and the number of "borderline" observations for the BOOT method),
- the number of false positives,
- the quality of the estimated EDR direction $d^2(E, \widehat{E}_\star)$ where $\widehat{E}_\star$ is the estimated EDR space based on the sample without the outliers (and the "borderline" observations) detected by the method $\star$,
- the MSE evaluated on the sample without the detected outliers (and the "borderline" observations for BOOT method).

To visualize and easily compare all these indicators, boxplots were used. According to results available on Fig. 5, all the three methods allow to reduce the distance to the true model. All methods, and even the model based on the complete dataset (in yellow), naturally perform better if the size of the available sample ($\tilde{n} + \tilde{\tilde{n}}$) increases. If the number of outliers is $\tilde{\tilde{n}} = 10$, the model based on the complete dataset shows poorer results whatever the number $\tilde{n}$. Note that, for a given number $\tilde{\tilde{n}}$ of outliers, the proportion of outliers naturally decreases as the sample size increases. BOOT seems to suffer from a large proportion of outliers only when the sample size is small.

Fig. 6 shows the MSE's for all the proposed methods for $p = 5$ and $\tilde{\tilde{n}} = 10$. Other simulations have been conducted and results are not provided because of

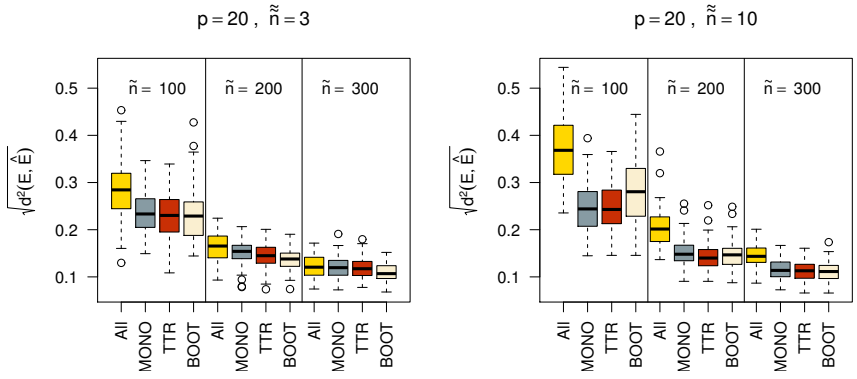

Fig. 5 Boxplots the quality measures of the estimated EDR space based on the $\sqrt{d^2(E, \widehat{E}_\star)}$ values, for simulated datasets with $p = 20$ and $\tilde{n} \in \{3, 10\}$. "All" stands for the estimation of the EDR using the complete sample.

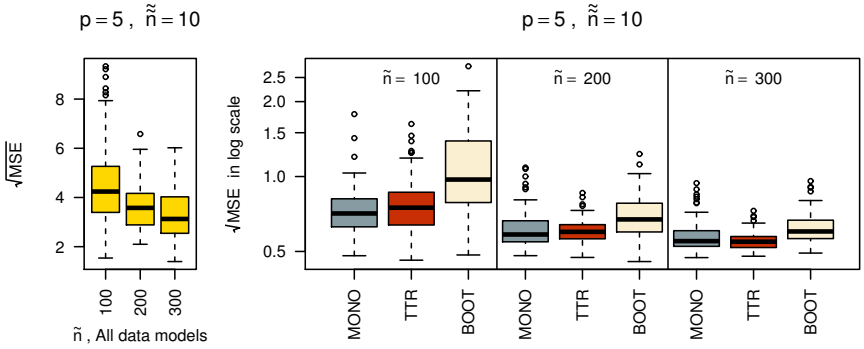

Fig. 6 Root MSE for the different methods. Visualizations of the errors for the complete dataset model (in yellow) and for the three outlier detection methods have been split in two graphics since scales are different. Here $p = 5$ and $\tilde{n} = 10$ have been detailed.

redundancy in the associated comments. Errors are larger for the complete dataset model (in yellow) than for any of the three methods but tend to decrease as $\tilde{n}$ increases and thus the proportion of outliers decreases. TTR seems to provide the best results for large sample sizes (and thus for low proportions of outliers), while BOOT shows larger errors, especially when $\tilde{n}$ is small (and thus when the proportion of outliers is high). An explanation of the phenomenon is that MSE is computed on the sample without the outliers. In that context, the MONO and TTR methods that select extreme (or "borderline") observations as outliers tend to get smaller MSE. On the contrary, the BOOT approach does not exclude these "borderline" observations which are more difficult to predict correctly, leading to a larger MSE. The MSE descriptor must be interpreted with this remark in mind, as well as by taking into account the number of false positives of each method, which is done thanks to Fig. 7.

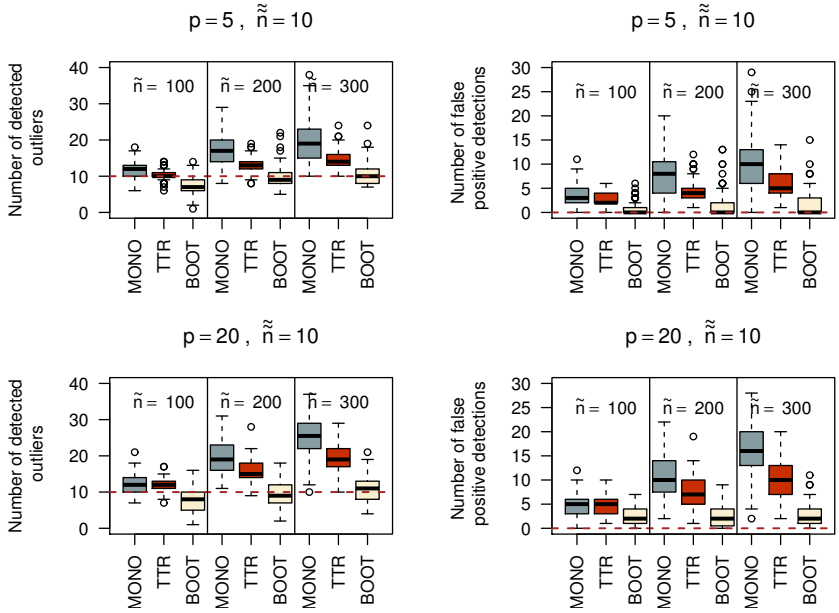

Fig. 7 Number of detected outliers (left column) and number of false positive detections (right column) for different values of p, $\tilde{n}$ and $\tilde{\tilde{n}} = 10$.

Whatever the sets of parameters in Fig. 7, BOOT is the only method that seems to be able to select the true outliers without selecting too many false positives (i.e. individuals detected as outliers when they are not). BOOT seems to be the most efficient method by showing the lowest number of false positives for all $\tilde{n}$. The number of false positives stays somewhat constant over the sample size $\tilde{n}$ for BOOT but increases with $\tilde{n}$ for the other two methods. MONO and TTR methods seem to have a sensibility to $\tilde{n}$ with an increase of the numbers of detected outliers and false positives as the sample size increases (and thus as the proportion of outliers decreases since their number $\tilde{\tilde{n}}$ is fixed at 10).

6 A real data application

Daily measurements of meteorological variables and ozone concentration are available in the dataset "ozone" (Source: [11]). More precisely, this dataset contains $n = 112$ daily measurements of meteorological variables (wind speed, temperature, rainfall, cloudiness) and ozone concentration recorded in Rennes (France) in summer 2001. In this study, an individual is a day. Eleven numerical variables are measured with no missing values:

- `maxO3`: maximum of daily ozone concentration measured in gr/m^3,

- `T9, T12, T15`: daily temperatures measured in degree Celsius at 9, 12 and 15h (called "temperature" variables hereafter),
- `Ne9, Ne12, Ne15`: cloudiness measured at 9, 12 and 15h (called "cloudiness" variables hereafter),
- `Vx9, Vx12, Vx15`: wind speed (E-W component) measured at 9, 12 and 15h (called "wind" variables hereafter),
- `maxO3v`: maximum concentration of ozone measured the day before.

The initial objective is to explain the maximum of daily ozone concentration (the response variable y is thus `maxO3`) by the $p = 10$ variables available (`T9, T12, T15, Ne9, Ne12, Ne15, Vx9, Vx12, Vx15, maxO3v`). Hereafter, let x be the vector of these ten covariates. To do this, the semiparametric regression model (2), is used and the EDR space $E = \text{Span}(\beta)$ is estimated by the usual SIR method (with the number of slices $H = 10$) while the link function f is estimated using the kernel estimator with the Gaussian kernel and the bandwidth tuned via cross-validation. Our aim is here to detect the presence or absence of outliers in this dataset. The proposed three outlier detection methods (MONO, TTR with $R = 1000$ and BOOT with $B = 1000$) are compared.

The naive MONO method does not detect outliers. The TTR method provides 9 outliers and the BOOT method identify 4 "borderline" observations and no outlier (see the corresponding plots in Fig. 8 respectively at the top left and at the top right). Among the 9 TTR's outliers, 4 of them are the BOOT's "borderline" observations. These 4 observations correspond to specific days in terms of road traffic, since these are days of major departures or returns from summer holidays in France. It is known that ozone pollution is also due to car traffic, but the built model is based only on weather data and does not take into account this important source of pollution. It is therefore quite natural that these 4 days correspond to individuals outside the model's standards. The 5 other specific TTR's outlier observations are closer to the scatterplot structure and they correspond to the days of early June, mid-June (music festival on the first day of summer), late July (end of a week) and mid September.

In order to improve the final model, the method introduced by [22] for selecting the relevant variables based on variable importance is now applied on the sample without the outliers (TTR method) or the "borderline" observations (BOOT method). Only the following $p^* = 4$ covariates are then selected: a temperature variable, `T12`, a cloudiness variable, `Ne9`, a wind variable, `Vx9` and the maximum concentration of ozone measured the day before, `maxO3v`. This is not surprising since the 3 variables of temperature (resp. cloudiness, wind speed) are strongly correlated with each other. The corresponding EDR directions are very close: $\hat{b}_{\text{SIR}}^{\text{TTR}} = (0.778, -0.565, 0.258, 0.094)'$ and $\hat{b}_{\text{SIR}}^{\text{BOOT}} = (0.660, -0.724, 0.175, 0.094)'$.

Finally, for the outlier detection method TTR (resp. BOOT), the plot of the estimated indices $x_i'\hat{b}_{\text{SIR}}^{\text{TTR}}$ (resp. $x_i'\hat{b}_{\text{SIR}}^{\text{BOOT}}$) versus the y_i's for the corresponding samples without outliers (resp. "borderline" observations) and the associated estimated link function (solid blue curve) are provided in Fig. 8 (at the bottom, on the right, resp. on the right). These two graphics are very similar and show an increasing link between the estimated index and the response variable `maxO3`. Then, it is possible to interpret the coefficients of the estimated EDR direction $\hat{b}_{\text{SIR}}^{\text{TTR}}$ (or similarly $\hat{b}_{\text{SIR}}^{\text{BOOT)}}$ using their

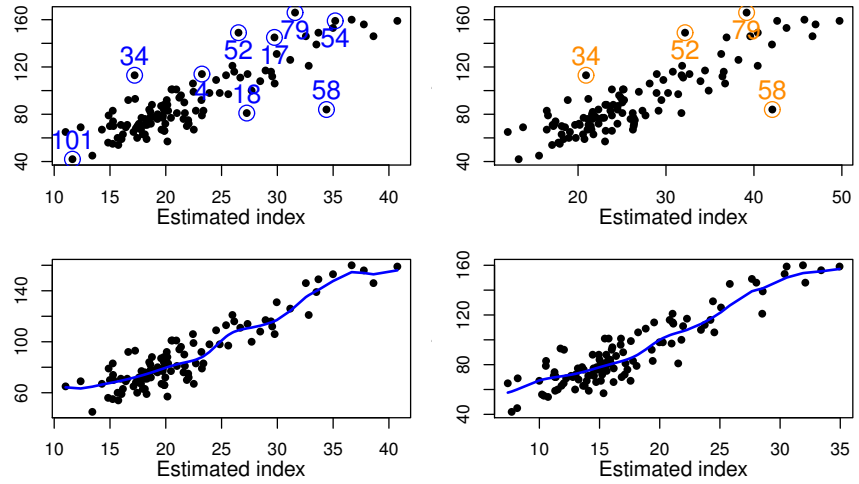

Fig. 8 Study of Ozone dataset. Top-Left: Selected outliers with TTR method. Top-Right: Selected "borderline" observations with BOOT method. Bottom-Left: plot of the y_i's (values of `max03`) versus the final estimated indices based on the $n^*_{\text{TTR}} = n - 9$ observations, i.e. removing the 9 selected outliers. Bottom-Right: plot of the y_i's (values of `max03`) versus the final estimated indices based on the $n^*_{\text{BOOT}} = n - 4$, removing the 4 selected "borderline" observations. The corresponding estimated link functions (solid blue curve) are superimposed on the last two plots.

signs. The variable `T12` (resp. `Vx9` and `max03v`) has a positive coefficient which means that an increase in daily temperatures at 12h (resp. of the wind speed at 9h, or of maximum concentration of ozone measured the day before) implies an increase of the estimated index and this then implies (not surprisingly) an increase of maximum of daily ozone concentration. On the contrary, the variable `Ne09` has a negative coefficient and then an increase of its values leads to a decrease in the maximum of daily ozone concentration, which is relevant from an air pollution point of view.

7 Concluding remarks and extensions

Three computational outlier detection approaches for sliced inverse regression have been presented. In this work, the original idea is to consider potential outliers that are outliers only in the SIR model and that are not detectable outliers by studying only their distribution in x or y. Thus considering the plot of the estimated indices versus the dependent variable, only outliers can appear in y. The case of outliers in x or in y are not considered here since the corresponding observations should be detectable as outliers in an early stage before the SIR modeling step, and the dataset should then be cleaned up accordingly. The MONO, TTR and BOOT approaches were implemented in R and the code is available on https://github.com/hlorenzo/outlierSIR.

The philosophy of these approaches does not rely neither on the SIR method used in the first estimation step nor on the nonparametric regression used in the second estimation step. For example, instead of the usual SIR method, it is possible to use the SIR-II, SIR_α or SAVE methods among others. Moreover, the proposed approaches are also easily generalizable to the multiple-index model framework, i.e. when the dimension of the EDR space is equal to $K > 1$. All SIR-related methods, as well as non-parametric regression methods (like multivariate kernels), work well in this framework . However the non-parametric regression methods might suffer from the well-known curse of dimensionality. Note that the choice of the dimension K of this EDR subspace should be then discussed. Finally, these outlier detection approaches can also be extended to a q-dimensional response variable y. Several authors developed SIR-based methods to estimate the EDR space that is common to the q components of the multivariate response variable, see for instance [4, 12, 28, 31, 35] among others. However, the concept of an outlier in this multivariate framework must be first clarified since it is not entirely natural.

Acknowledgements Jérôme Saracco would like to sincerely thank Prof. Christine Thomas-Agnan for having "brought back in his luggage" the SIR method in Toulouse in the early 90s. His first research focused on contributions to SIR (master's thesis, doctoral thesis, first articles in international journals). Thank you for all the discussions I had with Christine throughout my career on many scientific subjects (non-parametric estimation, conditional quantiles, etc.) and many other subjects.

References

1. Auger, I., Lawrence, C. (1989). Algorithms for the Optimal Identification of Segment Neighborhoods. *Bulletin of Mathematical Biology*, 51:1, 39–54.
2. Azais, R., Gegout-Petit, A., Saracco, J. (2012). Optimal quantization applied to Sliced Inverse Regression. *Journal of Statistical Planning and Inference*, 142, 481–492.
3. Babos, S., Artemiou, A. (2020). Sliced inverse median difference regression. *Stat Methods & Applications*.
4. Barreda, L., Gannoun, A., Saracco, J. (2007). Some extensions of multivariate sliced inverse regression. *Journal of Statistical Computation and Simulation*, 77:1, 1–17.
5. Cai, Z., Li, R., Zhu, L. (2020). Online Sufficient Dimension Reduction Through Sliced Inverse Regression. *Journal of Machine Learning Research*, 21:10, 1–25.
6. Chavent, M., Girard, S., Kuentz-Simonet, V., Liquet, B. Nguyen, T.M.N., Saracco, J. (2014). A sliced inverse regression approach for data stream. *Comput. Stat.*, 29, 1129–1152.
7. Chen, C.-H., Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8:2, 289–316.
8. Chiancone, A., Forbes, F., Girard, S. (2017). Student Sliced Inverse Regression. *Comput. Stat. Data Anal.*, 113, 441–456.
9. Cook, R. D. (2000). SAVE: A method for dimension reduction and graphics in regression. *Commun. Stat. - Theory Methods*, 29, 2109–2121.
10. Cook, R., Critchley, F. (2000). Identifying Regression Outliers and Mixtures Graphically. *J. Am. Stat. Assoc.*, 95(451), 781–794.
11. Cornillon, P.-A., Guyader, A., Husson, F., Jégou, N., Josse, J., Kloareg, M., Matzner-Lober, E., Rouvière, L. (2012). *R for Statistics*. Chapman & Hall/CRC, Rennes.
12. Coudret, R., Girard, S., Saracco, J. (2014). A new sliced inverse regression method for multivariate response. *Computational Statistics and Data Analysis*, 77 , 285–299.

13. Dikheel, T.R. (2014). Robust Sliced Inverse Regression. *J. Adm. and Eco. Sciences*, 15:1, 227–242.
14. Dong, Y., Yu, Z., Zhu, L. (2015). Robust inverse regression for dimension reduction. *J. Multivar. Anal.*, 134, 71–81.
15. Duan, N., Li, K. C. (1991). Slicing regression: a link-free regression method. *Ann. Stat.*, 19, 505–530.
16. Edwards, A., Cavalli-Sforza, L. (1965). Method for Cluster Analysis. *Biometrics*, 21:2, 362–375.
17. Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Am. Stat. Assoc.*, 93, 132–140.
18. Gannoun, A., Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, 13, 297–310.
19. Gather, U., Hilker, T., Becker, C. (2002). A note on outlier sensitivity of Sliced Inverse Regression. *Statistics*, 36:4, 271–281.
20. Hall, P., Li, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Stat.*, 21:2, 867–889.
21. Hsing, T. (1999). Nearest neighbor inverse regression. *Ann. Stat.*, 27:2, 697–731.
22. Jlassi, I., Saracco, J. (2019). Variable importance assessment in sliced inverse regression for variable selection. *Commun. Stat. - Simul. Comput.*, 48:1, 169–199.
23. Killick, R., Eckley, I.A., (2014). Changepoint: An R Package for Changepoint Analysis. *J. Stat. Softw.*, 58 (3).
24. Killick, R., Fearnhead, P., Eckley, I.A. (2012). Optimal Detection of Changepoints with a Linear Computational Cost. *J. Am. Stat. Assoc.*, 107(500), 1590–1598.
25. Li, B. (2018). *Sufficient dimension reduction. Methods and applications with R*. Chapman and Hall/CRC, New York.
26. Li, K.-C. (1991). Sliced inverse regression for dimension reduction, with dicussion. *J. Am. Stat. Assoc.*, 86, 316–342.
27. Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Am. Stat. Assoc.*, **87**, 1025-1039.
28. Li, K.-C., Aragon, Y., Shedden, K., Thomas Agnan, C. (2003). Dimension reduction for multivariate response data. *J. Am. Stat. Assoc.*, 98:461, 99–109.
29. Li, Y., Zhu, L. (2007). Asymptotics for sliced average variance estimation. *Ann. Stat.*, 35, 41–69.
30. Liquet, B., Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Commun. Stat. - Simul. Comput.*, 37:6, 1198–1218.
31. Lue, H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference*, 139:8, 2656–2664.
32. Prendergast, L.A. (2006). Detecting influential observations on the SIR e.d.r. space. *Aust. New Zeal. J. Statist.*, 48, 285–304.
33. Prendergast, L.A. (2007). Implications of influence function analysis for sliced inverse regression and sliced average variance estimation. *Biometrika*, 94:3, 585–601.
34. Saracco, J. (1997). An asymptotic theory for Sliced Inverse Regression. *Commun. Stat. - Theory Methods*, 26, 2141–2717.
35. Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *J. Multivar. Anal.*, 96, 117–135.
36. Schimek, M. G. (Ed.). (2013). *Smoothing and regression: approaches, computation, and application.*. John Wiley & Sons.
37. Singh, K., Xie, M. (2003). Bootlier-Plot: Bootstrap Based Outlier Detection Plot. *Sankhya, Ser. A*, 65:3, 532–559
38. Yin, X., Seymour, L. (2005). Asymptotic distributions for dimension reduction in the SIR-II method. *Statistica Sinica*, 15:4, 1069–1079.
39. Zhu, L.X., Miao, B., Peng, H. (2006). On sliced inverse regression with large dimensional covariates. *J. Am. Stat. Assoc.*, 101, 630–643.
40. Zhu, L., Zhu, L. (2007). On kernel method for sliced average variance estimation. *J. Multivar. Anal.*, 98, 970–991.