ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2019)
81, *Part* 2, *pp.* 385–408

Multiple influential point detection in high dimensional regression spaces

Junlong Zhao,

Beijing Normal University, People's Republic of China

Chao Liu and Lu Niu

Beihang University, People's Republic of China

and Chenlei Leng

University of Warwick, Coventry, and Alan Turing Institute, London, UK

[Received February 2017. Final revision December 2018]

Summary. Influence diagnosis is an integrated component of data analysis but has been severely underinvestigated in a high dimensional regression setting. One of the key challenges, even in a fixed dimensional setting, is how to deal with multiple influential points that give rise to masking and swamping effects. The paper proposes a novel group deletion procedure referred to as multiple influential point detection by studying two extreme statistics based on a marginal-correlation-based influence measure. Named the min- and max-statistics, they have complementary properties in that the max-statistic is effective for overcoming the masking effect whereas the min-statistic is useful for overcoming the swamping effect. Combining their strengths, we further propose an efficient algorithm that can detect influential points with a prespecified false discovery rate. The influential point detection procedure proposed is simple to implement and efficient to run and enjoys attractive theoretical properties. Its effectiveness is verified empirically via extensive simulation study and data analysis. An R package implementing the procedure is freely available.

Keywords: False discovery rate; Group deletion; High dimensional linear regression; Influential point detection; Masking and swamping; Robust statistics

1. Introduction

Recent decades have witnessed an explosion of high dimensional data in applied fields including biology, engineering, finance and many other areas. Given a data set consisting of $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ where $Y_i \in \mathbb{R}$ is the response and $\mathbf{X}_i \in \mathbb{R}^p$ is the covariate for the ith observation, the main interest is often to conduct a regression analysis to relate Y to $\mathbf{X}$, the simplest model for which takes the linear form.

A usual assumption in linear regression is that the observations are all generated from the same model. In many applications, however, the data that are collected often contain contaminated or noisy observations due to a plethora of reasons. Those observations exerting great influence on statistical analysis, thus named influential points, can seriously distort all aspects of data analysis such as altering the estimate of the regression coefficient and swaying the outcome of

Address for correspondence: Chenlei Leng, Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
E-mail: C.Leng@warwick.ac.uk

1369–7412/19/81385

statistical inference (Draper and Smith, 2014). Thus, when influential points are present, fitting the model based on a clean data assumption leads to at best a very crude approximation to the model and at worst a completely wrong solution. For fixed dimensional models, we refer the reader to Cook (1977), Belsley *et al*. (1980), Chatterjee and Hadi (1986), Imon (2005), Zhu *et al*. (2007, 2012) and Nurunnabi *et al*. (2014), among many others. For high dimensional models, Zhao *et al*. (2013) found that influential observations could negatively impact many methods that have recently been developed for dealing with high dimensionality, such as the lasso for variable selection (Tibshirani, 1996) and sure independence screening for variable screening (Fan and Lv, 2008).

As a result, influence diagnosis has been long recognized as a central problem in statistical analysis. An entire line of research has been devoted to devising robust methods that are less prone to influential observations; see, for example, the books on robust regression by Maronna *et al*. (2006), Huber and Ronchetti (2009) and Rousseeuw and Hubert (2011), as well as the papers by Wang *et al*. (2007) and Fan *et al*. (2014) for variable selection with heavy-tailed noise and She and Owen (2011) for an outlier robust method for the mean shift model. However, identifying influential points can often be of major scientific interest. For multivariate high dimensional data containing only $\mathbf{X}_i$s, Aggarwal and Yu (2001) proposed to use projection and Filzmoser *et al*. (2008) and Shieh and Hung (2009) applied principal component analysis, whereas Ro *et al*. (2015) used a robust covariance matrix estimator for defining distance.

When p is fixed, an attractive measure is to quantify individual observations' influence in changing the ordinary least squares (OLS) estimator and resulting quantities; see, notably, Cook's distance (Cook, 1977), Studentized residuals (Velleman and Welsch, 1981), DFFITS (Welsch and Kuh, 1977; Belsley *et al*., 1980) and Welsch's distance (Welsch, 1982). Since these measures are all based on OLS estimation, they are not applicable to high dimensional data. In contrast, the problem of influence diagnosis in high dimensional regression has received little attention, mainly due to the difficulty in establishing a coherent theoretical framework, even in a fixed dimension setting, and a lack of easily implementable procedures. To overcome this, Zhao *et al*. (2013) found that influence can be measured by examining how an individual observation affects the marginal correlation between the response and the predictors, which is a ubiquitous quantity in almost all aspects of regression analysis. They proposed a high dimensional influence measure named HIM to flag those points that have a great influence on the calculated value of marginal correlations as influential, in a sense to be defined later. An attractive feature of HIM is that its asymptotic properties can be rigorously established to enable the use of a multiple-testing procedure for detecting influential points.

However, similarly to many fixed dimensional measures, HIM is based on the idea of leave-one-out observation, i.e. to quantify the influence of an observation, one compares a predefined measure that is evaluated on the whole data set and the measure that is evaluated on a subset of the data leaving out the observation under investigation. Because of this, HIM is useful for detecting the presence of a single influential point. In practice, however, multiple influential observations are commonly encountered and it is not appropriate to apply a test for a single influential point sequentially to detect multiple points. However, detecting multiple influential observations is much more challenging, because of the notorious 'masking' and 'swamping' effects (Hadi and Simonoff, 1993). Specifically, masking occurs when an influential point is not detected as influential, whereas swamping occurs when a non-influential point is classified as influential. In the language of multiple testing, masking is the problem of obtaining false negative results and swamping is the problem of obtaining false positive results. To handle these effects, group-deletion-based methods have been proposed for fixed dimensional problems (Rousseeuw

and van Zomeren, 1990; Hadi and Simonoff, 1993; Imon, 2005; Pan *et al.*, 2000; Nurunnabi *et al.*, 2014; Roberts *et al.*, 2015) but it is currently an open problem for high dimensional problems.

The main aim of this paper is to propose a new procedure for detecting multiple influential points for high dimensional data based on HIM. Via random group deletion, we propose a novel procedure named MIP, short for multiple influential point detection for high dimensional data. Along with the process, we propose two novel quantities named max- and min-statistics to assess the extremeness of each point when data are subsampled. Our theoretical studies show that these two statistics have complementary properties. The min-statistic is useful for overcoming the swamping effect but less effective for masked influential observations, whereas the max-statistic is well suited for detecting masked influential observations but is less effective in handling the swamping effect. Combining their advantages, we propose a computationally simple algorithm for obtaining a clean subset of the data that contains no points that greatly influence the marginal correlation with high probability. This clean set of data is then served as the benchmark for assessing the influence of other observations, which permits us to control the false discovery rate (FDR) of influential points by using, for example, the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). Remarkably, the theoretical properties of max- and min-statistics can be studied and are rigorously established in this paper. We must point out that, even for fixed dimensional problems, there is a general lack of principled procedures for declaring significance for any defined influence measure. On the contrary, our proposed MIP procedure is the first theoretically justified method for the more challenging high dimensional regression setting.

Before we proceed, we highlight the usefulness of the max- and min-statistics via an analysis of the microarray data in Section 5. Fig. 1 plots the logarithms of the p-values that are associated with the max-statistic in Fig. 1(a) and the min-statistic in Fig. 1(b) of the observations. With a prespecified FDR of 0.05, using the min-statistic, we identify a set of seven influential observations, represented as the full circles in Figs 1(a) and 1(b). It is interesting that the MIP procedure combining the strengths of the two statistics identifies the same set of seven influential points. In contrast, using the max-statistic, four additional observations, represented as triangles in Fig. 1(a), are declared influential. These findings are consistent with our theory that the max-statistic tends to identify more influential observations, making it more suitable for overcoming the masking effect, but may suffer from the swamping effect. However, the fact that the min-statistic gives the same set of influential points as MIP in Fig. 1(b) implies that there may not be any masking effect in these data. Further analysis in Section 5 shows that the reduced data, which are obtained by removing the influential observations that are identified by MIP, result in a sparser model with a better fit, when the lasso is applied.

The rest of the paper is organized as follows. In Section 2, we review the high dimensional influence measure in Zhao *et al.* (2013). In Section 3, on the basis of the idea of random group deletion or leave-many-out observations, we propose max- and min-statistics for assessing extremeness and establish their theoretical properties. We show in theorem 1 that, surprisingly, when there is no influential point, these two statistics both follow a $\chi^2(1)$ distribution. When there are influential points, theorem 2 and theorem 3 show that, for a non-influential point, its max- and min-statistics still follow a $\chi^2(1)$ distribution. Furthermore with the presence of influential points, theorems 2 and 3 demonstrate that, under suitable conditions, the max- and min-statistics can identify the influential points with large probability. We then argue that these two statistics are complementary in detecting influential observations and we develop an algorithm to combine their strengths. Simulation results are presented in Section 4 and data analysis is provided in Section 5. In Section 6, we provide further discussion. All the proofs as well as

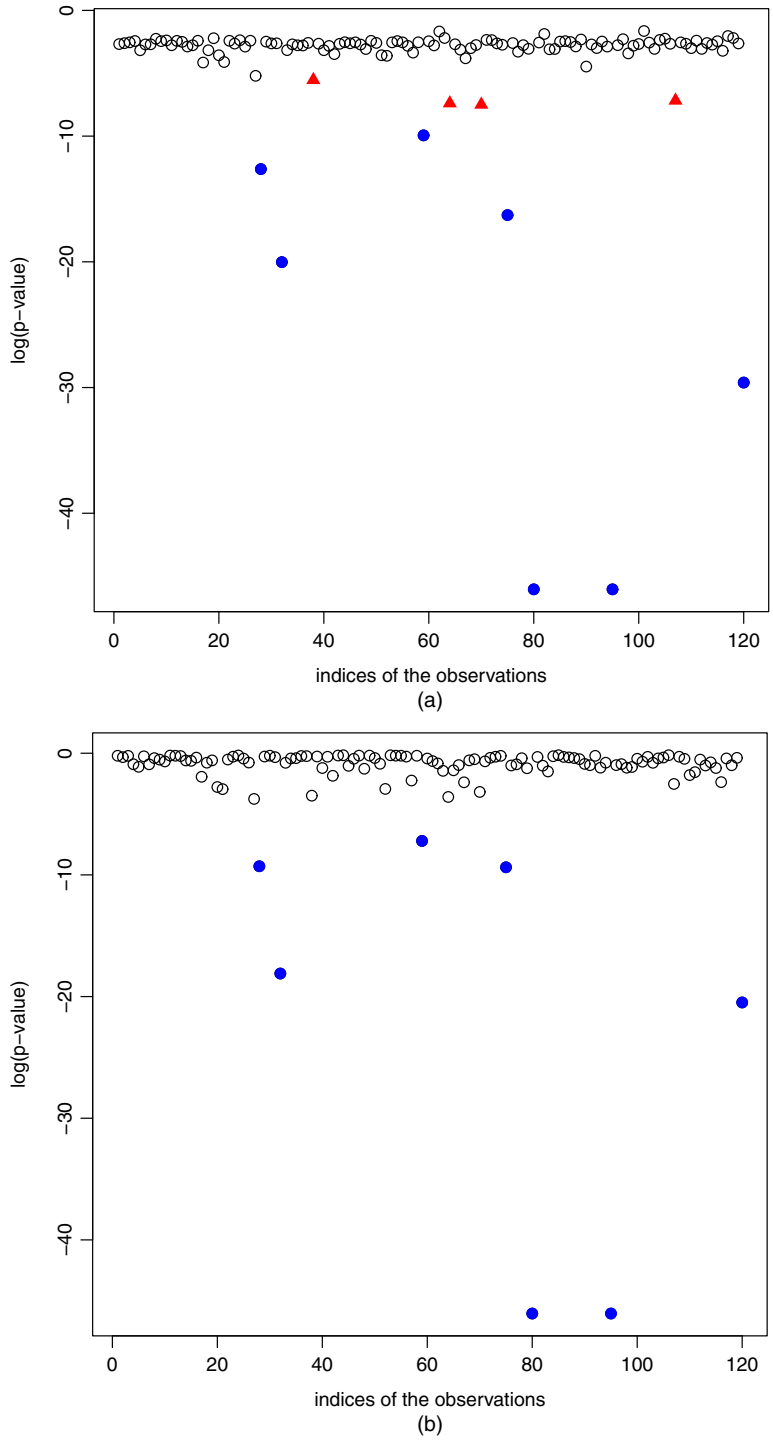

Fig. 1. Influential point detection by using (a) the max- or (b) min-statistic: in (a), identified influential points are denoted by either full circles or triangles, whereas, in (b), identified influential points are denoted by full circles; MIP identifies the seven as influential

the details of the simulation study are relegated to the on-line supplementary materials. An R package implementing MIP is freely available from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`

Here is the notation that is used throughout the paper. For any set A, we write $|A|$ and N_A interchangeably as its cardinality. Let S_{inf} and $S_{\text{inf}}^{\text{c}}$ be the set of the influential and non-influential observations respectively, such that $S_{\text{inf}} \cup S_{\text{inf}}^{\text{c}} = \{1, \ldots, n\}$. We denote $|S_{\text{inf}}| = n_{\text{inf}}$ as the size of the influential point set and $|S_{\text{inf}}^{\text{c}}| = n - n_{\text{inf}}$ as the number of non-influential points. Denote by $\|v\|$ the l_2-norm of a vector $v \in \mathbb{R}^m$. For any matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, $\|A\|$ denotes its spectral norm. Finally, let $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ and use C to denote a generic constant that may change depending on the context.

2. High dimensional influence measure, masking and swamping

Because HIM (Zhao *et al.*, 2013) is the influence measure that is used for our influence diagnosis, we first give a brief review. Assume that the non-influential observations are independent and identically distributed from the model

$$Y_i = \mathbf{X}_i^{\mathrm{T}} \beta + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $Y_i \in \mathbb{R}$ is the response variable, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}} \in \mathbb{R}^p$ is the p-dimensional predictor, $\beta \in \mathbb{R}^p$ is the coefficient and $\varepsilon_i \in \mathbb{R}$ is normally distributed random noise with $\operatorname{cov}(\mathbf{X}_i, \varepsilon_i) = 0$. Denote $\mu_y = E(Y_i)$, $\sigma_y = \operatorname{var}(Y_i)^{1/2}$ and $\mu_x = (\mu_{x1}, \ldots, \mu_{xp})^{\mathrm{T}} = E(\mathbf{X}_i)$ and $\sigma_{xj} = \operatorname{var}(X_{ij})^{1/2}$, $1 \leqslant j \leqslant p$.

HIM defines the influence of a point by measuring its contribution to the average marginal correlation between the response and the predictors. Specifically, define $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)^{\mathrm{T}}$ where $\rho_j = \operatorname{corr}(X_{ij}, Y_i)$ is the marginal correlation between the jth variable and the response. From the data, we can obtain a sample estimate as $\hat{\rho}_j = \Sigma_{i=1}^n (X_{ij} - \hat{\mu}_{xj})(Y_i - \hat{\mu}_y)/(n\hat{\sigma}_{xj}\hat{\sigma}_y)$, for $j = 1, \ldots, p$, where $\hat{\mu}_{xj}$, $\hat{\mu}_y$, $\hat{\sigma}_{xj}$ and $\hat{\sigma}_y$ are the sample estimates of μ_{xj}, μ_y, σ_{xj} and σ_y respectively. The sample marginal correlation with the kth observation removed is similarly defined as $\hat{\rho}_j^{(k)}$ for $1 \leqslant k \leqslant n$. HIM then measures the influence of the kth observation by comparing the sample correlations with and without this observation, defined formally as

$$\mathbb{D}_k = p^{-1} \sum_{j=1}^{p} (\hat{\rho}_j - \hat{\rho}_j^{(k)})^2, \qquad 1 \leqslant k \leqslant n.$$

Intuitively, the larger $\mathbb{D}_k$ is, the more influential the corresponding observation is. When there is no influential point and $\min\{n, p\} \to \infty$, under mild conditions, it is proved that $n^2 \mathbb{D}_k \to_{\text{d}} \chi^2(1)$, where $\chi^2(1)$ is the χ^2-distribution with 1 degree of freedom. Based on this, we can formulate the problem of influential point detection as a multiple-hypothesis-testing problem where one tests n hypotheses, one for each observation, stating that the observation under investigation is non-influential. Subsequently, the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) for multiple testing can be used to control the false discovery rate.

Since HIM is based on the leave-one-out idea, the $\chi^2(1)$ distribution derived is invalid whenever there are one or more influential points, i.e., for a non-influential point, the presence of even one single influential point can distort the null distribution of its HIM value according to the definition above. Similarly, the presence of more than one influential point can distort the HIM value of an influential point as well. This is the manifestation of a more general difficulty of multiple-influential-point detection where the masking and swamping effects greatly hinder the

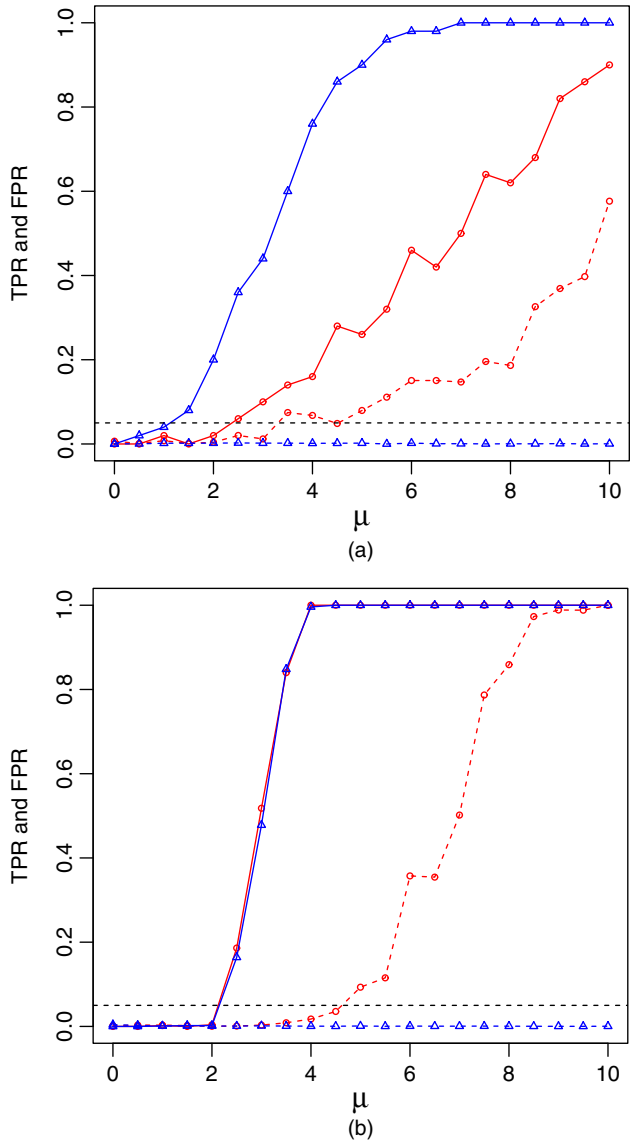

Fig. 2. Performance comparison between HIM and MIP with $(n, p) = (100, 1000)$ (– – –, nominal FPR, set at $\alpha = 0.05$; —○—, TPR, HIM; - -○- -, FPR, HIM; —△—, TPR, MIP; - -△- -, FPR, MIP): (a) masking effect example (example 1); (b) swamping effect example (example 2)

usefulness of any leave-one-out procedures. To appreciate how masking and swamping effects negatively impact the performance of HIM, we quickly look at examples 1 and 2 in Section 4. The data are generated such that there is a strong masking effect in example 1 and a strong swamping effect in example 2. The magnitude of these effects depends on a parameter denoted as μ. Fig. 2 presents a comparison of HIM in Zhao *et al.* (2013) and MIP proposed in this paper for detecting influence, when the nominal level that is used for declaring influential in the Benjamini–Hochberg procedure is set at $\alpha = 0.05$.

From Fig. 2(a), we see that the true positive rates (TPRs) of HIM are much lower than those

of MIP, i.e. HIM identifies much fewer influential points as influential and thus suffers severely from the masking effect. Meanwhile, the false positive rates (FPRs) of HIM are also much larger than the nominal level $\alpha = 0.05$ especially when μ becomes large, i.e. HIM identifies many more non-influential points as influential, meaning that HIM also suffers from the swamping effect. From Fig. 2(b), we see that HIM suffers from the swamping effect greatly, as the FPRs can be very close to 1 for large μ. In contrast, for both examples, the FPRs of the MIP procedure are controlled well below the nominal level whereas its TPRs are monotone functions of μ and eventually become 1 for large μ.

3. A random group deletion procedure

As discussed earlier, any measure based on the leave-one-out approach may be ineffective when there are multiple influential observations due to the masking and swamping effects. Since the number of influential observations is generally unknown in practice, it is natural to employ a notion of leave-many-out or group deletion which has been used for fixed dimensional problems in identifying multiple influential points (Lawrence, 1995; Imon, 2005; Nurunnabi, 2011; Nurunnabi *et al.*, 2014; Roberts *et al.*, 2015), where deletion is often made according to the magnitude of (Studentized) residuals or similar criteria with a good estimate of β necessary.

For our random group deletion procedure, we do something similar by choosing the subsets uniformly at random with replacement. Thus, the marginal correlations based on these subsets can be seen as some kind of perturbation to the marginal correlations based on the whole sample. It turns out that their extremeness can be summarized by two extremal statistics whose theoretical properties can be studied, as we do below. Existing group deletion procedures are not employed in a similar way and they do not usually give theoretically tractable results. Since the emphasis of this paper is on influence diagnosis, throughout the paper we assume that the non-influential observations are generated from model (2.1) whereas the influential observations are from a different model, the properties of which will be specified later.

Write $Z_k = (\mathbf{X}_k, Y_k), 1 \leqslant k \leqslant n$, as the kth data point. For any fixed k, to check whether Z_k is influential or not, we draw with replacement some subsets $A_1, \ldots, A_m \subset \{1, \ldots, n\}/\{k\}$ uniformly at random, i.e. these subsets do not include Z_k. The choice of m will be discussed in Section 4. Write $|A_r| = n_{\text{sub}} - 1$ where $n_{\text{sub}} = k_{\text{sub}} n + 1$ for some $k_{\text{sub}} \in (0, 1)$. We make the following assumption on n_{inf} and k_{sub}.

Assumption 1. Denote $\delta_{\text{inf},n} = n_{\text{inf}}/n$ which is allowed to vary with n. Assume that $0 \leqslant \delta_{\text{inf},n} < \frac{1}{2} - \delta_1$ for some $\delta_1 > 0$ independent of n. We take $k_{\text{sub}} > \limsup_n \delta_{\text{inf},n} + \delta_1$.

Assumption 1 allows $\min_n \delta_{\text{inf},n} \to 0$. Without loss of generality, from now on, we take a conservative choice $k_{\text{sub}} = \frac{1}{2}$ as non-influential points are expected to outnumber the influential points. For $1 \leqslant r \leqslant m$, let B_r be the subset of non-influential observations in A_r and denote its size as $N_{B_r} = |B_r|$. Under assumption 1, we have $\min_{1 \leqslant r \leqslant m} N_{B_r} > \delta_1 n$, i.e., for any subset A_r, the number of non-influential observations does not vanish.

For $1 \leqslant r \leqslant m$, let $A_r^{(+k)} = A_r \cup \{k\}$ which is of size n_{sub}. For Z_k, we compute its influence measure with respect to the rth random subset A_r as

$$\mathcal{D}_{r,k} = p^{-1} \|\hat{\boldsymbol{\rho}}_{A_r^{(+k)}} - \hat{\boldsymbol{\rho}}_{A_r}\|^2, \qquad 1 \leqslant r \leqslant m,$$

where $\hat{\boldsymbol{\rho}}_{A_r}$ and $\hat{\boldsymbol{\rho}}_{A_r^{(+k)}}$ denote the estimate of $\boldsymbol{\rho}$ based on observations in A_r and $A_r^{(+k)}$ respectively. We are now ready to define the following two extreme statistics:

$$T_{\min,k,m} = \min_{1\leqslant r\leqslant m} n_{\text{sub}}^2 \mathcal{D}_{r,k},$$

$$T_{\max,k,m} = \max_{1\leqslant r\leqslant m} n_{\text{sub}}^2 \mathcal{D}_{r,k}.$$

We name them the min- and max-statistic respectively as they measure the extremeness of the influence measures based on randomly sampled data. Note that $T_{\min,k,m}$ and $T_{\max,k,m}$ are functions of $\mathbb{Z}_n = \{Z_k, 1 \leqslant k \leqslant n\}$ and $\mathbb{A}_m = \{A_r, 1 \leqslant r \leqslant m\}$. The dependence on $\mathbb{Z}_n$ and $\mathbb{A}_m$ is summarized by using a subscript m to simplify the notation. These two statistics are invariant to the rotation of the covariates and to the scale translation of the response. Let

$$\mathcal{B} = \{B : B \subset S_{\text{inf}}^{\text{c}}, N_B \geqslant n\delta_1\}.$$

To establish the asymptotics of $T_{\min,k,m}$ and $T_{\max,k,m}$, we first characterize a key quantity

$$J_{\max,n} = \max_{B\in\mathcal{B}} J_B,$$

where J_B is defined as

$$J_B = p^{-1} \sum_{j=1}^{p} \left(\frac{1}{N_B} \sum_{t\in B} \hat{Y}_t \hat{X}_{tj} \right)^2 = p^{-1} \left\| \frac{1}{N_B} \sum_{t\in B} \hat{Y}_t \hat{\mathbf{X}}_t \right\|^2,$$

with $\hat{Y}_t = \hat{\sigma}_y^{-1}(Y_t - \hat{\mu}_y)$, $\hat{\mathbf{X}}_t = \hat{D}_x^{-1}(\mathbf{X}_t - \hat{\mu}_x)$, $1 \leqslant t \leqslant n$, and $\hat{D}_x$ being the estimate of $D_x = \text{diag}(\sigma_{x1}, \ldots, \sigma_{xp})$, a diagonal matrix in $\mathbb{R}^{p\times p}$. By definition, J_B is the square of the l_2-norm associated with the non-influential observations and is therefore unknown. Denote $\dot{\mathbf{X}}_t = D_x^{-1}(\mathbf{X} - \mu_x)$ as the population version of $\hat{\mathbf{X}}_t$ and note that $\dot{Y}_t$ is the population version of $\hat{Y}_t$. Without loss of generality, we assume in model (2.1) that $\mu_y = \mu_x = 0$ and $\sigma_y = \sigma_{xj} = 1$, $1 \leqslant j \leqslant p$. Moreover, we make the following assumptions.

Assumption 2. For $1 \leqslant j \leqslant p$, ρ_j is constant and does not change as p increases.

Assumption 3. For the covariance matrix of the covariates $\Sigma = \text{cov}(\mathbf{X}_i)$ with eigendecomposition $\Sigma = \Sigma_{j=1}^{p} \lambda_j u_j u_j^{\mathrm{T}}$, we assume that $l_p = \Sigma_{j=1}^{p} \lambda_j^2 = O(p^r)$ for some $0 \leqslant r < 2$.

Assumption 4. The predictor $\mathbf{X}_i$ follows a multivariate normal distribution and the random noise $\varepsilon_i \in \mathbb{R}$ follows a normal distribution with mean 0 and an unknown variance.

Assumption 5. Let $(Q_y, R_y) = ((\hat{\mu}_y - \mu_y)/\sigma_y, \sigma_y/\hat{\sigma}_y - 1)$, $S_{Qy} = \limsup_{n\to\infty} E(n^{1/2} Q_y)^8$ and $S_{Ry} = \limsup_{n\to\infty} E(n^{1/2} R_y)^8$. Assume that S_{Q_y} and S_{R_y} are finite. Furthermore, there are constants $0 < K, C < \infty$, independent of n and p, such that, for any $t > 0$,

$$\max_{1\leqslant j\leqslant p} P(|\hat{\mu}_{xj} - \mu_{xj}| > t/\sqrt{n}) \leqslant C \exp(-t^2/K),$$

$$\max_{1\leqslant j\leqslant p} P(|\hat{\sigma}_{xj}/\sigma_{xj} - 1| > t/\sqrt{n}) \leqslant C \exp\{-\min(t/K, t^2/K^2)\}.$$

Assumptions 2–4 on the non-influential observations were also made in Zhao *et al.* (2013). Since it is assumed that $\sigma_{xj} = 1$, $1 \leqslant j \leqslant p$, we have $\text{tr}(\boldsymbol{\Sigma}) = p$ and consequently it holds that $l_p \leqslant p^2$ by the Cauchy–Schwarz inequality. When $l_p = p^2$, $\boldsymbol{\Sigma}$ is a degenerate matrix with rank 1 and assumption 3 rules out this case. In contrast, assumption 3 applies when the largest eigenvalue of $\boldsymbol{\Sigma}$ is bounded. Assumption 5 is similar to but stronger than condition (C.4) of Zhao *et al.* (2013), where only eighth moments of $n^{1/2}(\hat{\mu}_{xj} - \mu_{xj})$ and $n^{1/2}(\hat{\sigma}_x/\sigma_x - 1)$ are required. In

assumption 5, $n^{1/2}(\hat{\mu}_{xj} - \mu_{xj})$ is assumed to have sub-Gaussian tails and $n^{1/2}(\hat{\sigma}_{xj}/\sigma_{xj} - 1)$s have subexponential tails. This assumption is satisfied for the sample mean and the sample variance under the normality of $(\mathbf{X}_i, Y_i)$s. As alternatives to the sample estimates, robust estimates of μ_x, μ_y, σ_{xj} and σ_y can also be used in practice. For example, we can estimate μ_{xj} and μ_y by the sample median and σ_{xj} by the median absolute deviation estimator. These estimates satisfy assumption 5 by noting the normality of $(\mathbf{X}_i, Y_i)$s. These robust estimates are the quantities that are used in our numerical examples.

We now quantify the magnitude of $J_{\max,n}$, the maximum effect of the non-influential points, which is independent of m and $\mathbb{A}_m$ and is a key quantity for establishing the asymptotic properties of the min- and max-statistic.

Lemma 1. Assume that the non-influential observations, generated from model (2.1), satisfy assumptions 2–4 and 5. Assume further that $\mathcal{B} \neq \emptyset$ and that $\xi_{n,p} = n^{-1/2} \log(p) \log(n) \log(np) \to 0$. Then

$$J_{\max,n} = O_p(\xi_{n,p} + p^{-1} l_p^{1/2}).$$

The conclusion is derived by bounding the maximum of $p^{-1} \| \hat{Y}_t \hat{\mathbf{X}}_t \|^2$ over t. Obviously, $\xi_{n,p} \to 0$ if $n^{-1/4+\epsilon_0} \log(p) \to 0$ for some sufficiently small $\epsilon_0 > 0$. Under assumption 1, it is obvious that $N_{B_r} \geqslant n\delta_1$ and consequently $B_r \in \mathcal{B}$ for any $1 \leqslant r \leqslant \infty$. Then lemma 1 implies immediately that, under assumption 1,

$$\sup_{1 \leqslant r \leqslant \infty} J_{B_r} \leqslant J_{\max,n} = O_p(\xi_{n,p} + p^{-1} l_p^{1/2}).$$

Replacing $(\hat{\mathbf{X}}_t, \hat{Y}_t)$ by $(\dot{\mathbf{X}}_t, \dot{Y}_t)$ in J_B, it is shown in the proof that the corresponding statistic is no more than $O_p(n^{-1} + p^{-1} l_p^{1/2})$. On the basis of lemma 1, we make the following claim.

Theorem 1. Suppose that all observations are non-influential. Under assumption 1 and the assumptions of lemma 1, for any $1 \leqslant k \leqslant n$, $T_{\min,k,m} \to_{\mathrm{d}} \chi^2(1)$ and $T_{\max,k,m} \to_{\mathrm{d}} \chi^2(1)$ uniformly over m and $\mathbb{A}_m$, which is denoted as over $(m, \mathbb{A}_m)$ for brevity, as $\min\{n, p\} \to \infty$.

Theorem 1 seems surprising at first glance, since we always have $T_{\min,k,m} \leqslant T_{\max,k,m}$. An explanation is in place. As we show below, $\mathcal{D}_{r,k}$ can be decomposed into two parts. The first part, depending on the quantity E_k to be defined soon, represents the effect of the observation Z_k, and the second part is controlled by $J_{\max,n}$. Since $J_{\max,n} = o_p(1)$ by lemma 1, the asymptotic distributions of $T_{\min,k,m}$ and $T_{\max,k,m}$ are mainly determined by E_k. Thanks to the blessing of dimensionality, we can show that E_k asymptotically has a $\chi^2(1)$ distribution. From theorem 1, when $T_{\max,k,m}$ or $T_{\min,k,m}$ is larger than $\chi^2_{1-\alpha}(1)$, the $100(1-\alpha)\%$ quantile of the $\chi^2(1)$ distribution, for some prespecified α such as 0.05, we declare that there are outliers.

Recall that B_r collects the indices of the non-influential observations in A_r. Let $O_r = A_r \setminus B_r$ be its complement in A_r. For each $1 \leqslant r \leqslant m$, it is obvious that $O_r \subseteq S_{\inf} \setminus \{k\}$, the latter equal to $S_{\inf}$ if $k \in S_{\inf}^{\mathrm{c}}$. Since $|A_r| = k_{\mathrm{sub}} n$, similarly to the proof of theorem 1, we have

$$\begin{aligned} n_{\mathrm{sub}}^2 \mathcal{D}_{r,k} &= p^{-1} \left\| \frac{1}{n_{\mathrm{sub}} - 1} \sum_{t \neq k, t \in A_r} \hat{Y}_t \hat{\mathbf{X}}_t - \hat{Y}_k \hat{\mathbf{X}}_k \right\|^2 \\ &= p^{-1} \left\| \frac{1}{n k_{\mathrm{sub}}} \sum_{t \in B_r} \hat{Y}_t \hat{\mathbf{X}}_t + \frac{1}{n k_{\mathrm{sub}}} \sum_{t \in O_r} \hat{Y}_t \hat{\mathbf{X}}_t - \hat{Y}_k \hat{\mathbf{X}}_k \right\|^2 \end{aligned}$$

$$:= p^{-1}\|W_{\text{non},k,B_r} + W_{\text{inf},k,O_r} - \hat{Y}_k\hat{\mathbf{X}}_k\|^2, \tag{3.1}$$

where $W_{\text{inf},k,O_r} = \Sigma_{t\in O_r}\hat{Y}_t\hat{\mathbf{X}}_t/nk_{\text{sub}}$ and $W_{\text{non},k,B_r} = \Sigma_{t\in B_r}\hat{Y}_t\hat{\mathbf{X}}_t/nk_{\text{sub}}$ are associated with influential and non-influential observations respectively. Under assumption 1 and the conditions in lemma 1, we see from lemma 1 that

$$\max_{1\leqslant r\leqslant\infty} p^{-1}\|W_{\text{non},k,B_r}\|^2 = \max_{1\leqslant r\leqslant\infty}\frac{N_{B_r}^2}{(nk_{\text{sub}})^2}J_{B_r} \leqslant \max_{1\leqslant r\leqslant\infty} J_{B_r} \leqslant J_{\max,n} = O_p(\xi_{n,p} + p^{-1}l_p^{1/2}). \tag{3.2}$$

Based on expressions (3.1) and (3.2), it is shown in lemma 1.3 of the on-line supplementary materials that, uniformly over $(m, k, \mathbb{A}_m)$, we have

$$\left|T_{\max,k,m} - \max_{1\leqslant r\leqslant m}(p^{-1}\|W_{\text{inf},k,O_r} - \hat{Y}_k\hat{\mathbf{X}}_k\|^2)\right| = o_p(1),$$

$$\left|T_{\min,k,m} - \min_{1\leqslant r\leqslant m}(p^{-1}\|W_{\text{inf},k,O_r} - \hat{Y}_k\hat{\mathbf{X}}_k\|^2)\right| = o_p(1).$$

Define

$$E_k = p^{-1}\|\hat{Y}_k\hat{\mathbf{X}}_k\|^2,$$

which represents the effect of the kth observation Z_k. Let $\mathbb{Z}_n^{\text{inf}} = \{Z_k, k\in S_{\text{inf}}\}$ be the influential observations and $\mathbb{O}_m = \{O_r, 1\leqslant r\leqslant m\}$ be the indices of the influential observations in the subsets. Define

$$F_{\min,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m) = \min_{1\leqslant r\leqslant m} p^{-1}\|W_{\text{inf},k,O_r}\|^2$$

and

$$F_{\max,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m) = \max_{1\leqslant r\leqslant m} p^{-1}\|W_{\text{inf},k,O_r}\|^2,$$

which quantify the maximum and minimum joint effects of the influential observations respectively. Thus the asymptotic behaviour of $T_{\max,k,m}$ and $T_{\min,k,m}$ mainly depends on the magnitude of E_k, $F_{\min,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m)$ and $F_{\max,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m)$.

Note that $F_{\min,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m)$ is a decreasing function of m, whereas $F_{\max,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m)$ is an increasing function of m. In lemma 1.1 in the on-line supplementary materials, we show that, uniformly over all $(k, m, \mathbb{O}_m)$,

$$F_{\max,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m) \leqslant \tilde{F}_{\max,k}(\mathbb{Z}_n^{\text{inf}}) \leqslant R_{\text{inf},n}^2 d_{S_{\text{inf}}}, \tag{3.3}$$

where

$$d_{S_{\text{inf}}} = \max_{t\in S_{\text{inf}}} E_t,$$

$$R_{\text{inf},n} = \delta_{\text{inf},n}/k_{\text{sub}}$$

and

$$\tilde{F}_{\max,k}(\mathbb{Z}_n^{\text{inf}}) = \max_{O\subset S_{\text{inf}}\setminus\{k\}} p^{-1}\|W_{\text{inf},k,O}\|^2.$$

Note that $\tilde{F}_{\max,k}(\mathbb{Z}_n^{\text{inf}})$ is independent of m and $\mathbb{O}_m$.

Deriving the upper bound of $F_{\min,k}(\mathbb{Z}_n^{\text{inf}}, \mathbb{O}_m)$ is rather involved and we present the result in lemma 1.2 of the on-line supplementary materials. We remark that the magnitude of

$F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ depends on the distribution of the influential observations, and it is generally difficult to quantify explicitly how large m should be. This is illustrated by two cases that are considered in lemma 1.2 of the supplementary materials. In the first case where the effect of influential observations is small, it holds that $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_p 0$ in probability uniformly over m, which means that m can be fixed. For the second case where m is sufficiently large (e.g. $mv_{\inf}^n \log(n)^{-1} \to \infty$ with $v_{\inf}$ defined in lemma 1.2 in the supplementary materials), we have $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_p 0$ uniformly over $\mathbb{Z}_n^{\inf}$, which means that $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ can be small as long as m is sufficiently large, regardless of the distribution of influential observations. Simulation results show that a small or moderate m is usually enough to obtain satisfactory empirical results. If m must be chosen judiciously, we have also proposed a numerical criterion as in Section 4, which works well in simulations. Below we state the properties of $T_{\max,k,m}$ and $T_{\min,k,m}$ separately in theorem 2 in Section 3.1 and theorem 3 in Section 3.2.

3.1. Max-statistic $T_{\max,k,m}$ for the kth point

In theorem 1, we have derived the null distribution of $T_{\max,k,m}$ and $T_{\min,k,m}$ when there is no influential point. We now study $T_{\max,k,m}$ when there are influential observations and develop the corresponding detection procedure. We have the following results.

Theorem 2. Under assumption 1 and the assumptions of lemma 1, when there are influential observations, the following two conclusions hold.

(a) Suppose further that $\tilde{F}_{\max,k}(\mathbb{Z}_n^{\inf}) \to_p 0$, as $\min\{n, p\} \to \infty$. If observation k is non-influential, i.e. $k \in S_{\inf}^c$, then $T_{\min,k,m}$ and $T_{\max,k,m}$ converge to $\chi^2(1)$ in distribution uniformly over $(m, \mathbb{A}_m)$, as $\min\{n, p\} \to \infty$.

(b) Suppose that Z_k is an influential point (i.e. $k \in S_{\inf}$). For $\epsilon_0 > 0$ sufficiently small, define

$$\mathcal{I}_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) = \{E_k^{1/2} > \chi^2_{1-\alpha}(1)^{1/2} + F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)^{1/2} + \epsilon_0\}.$$

Then, as $\min\{n, p\} \to \infty$, it holds that, uniformly over $(m, \mathbb{A}_m)$,

$$P\{T_{\max,k,m} > \chi^2_{1-\alpha}(1) | \mathbb{A}_m\} - P\{\mathcal{I}_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) | \mathbb{A}_m\} \geqslant 0.$$

Therefore, for any $m = m(n)$, if the following max-unmask condition holds,

$$P\{\mathcal{I}_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)\} \to 1, \qquad \text{as } \min\{n, p\} \to \infty,$$

then $P\{T_{\max,k,m} > \chi^2_{1-\alpha}(1)\} \to 1$, as $\min\{n, p\} \to \infty$.

The max-unmask condition becomes weaker as m increases, since $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ decreases with respect to m. Statement (b) of theorem 2 can be specified further by combining the properties of $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$. For example, by expression (2) of lemma 1.2 in the on-line supplementary materials, if $\delta_{\inf,n} = o(1)$ and $d_{S_{\inf}} = O_p(1)$, then $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_p 0$ uniformly over $(k, m, \mathbb{O}_m)$, as $\min\{n, p\} \to \infty$. In this case, the max-unmask condition can be updated as $P\{E_k > \chi^2_{1-\alpha}(1) + \epsilon_0\} \to 1$, as $\min\{n, p\} \to \infty$.

Under the condition in result (b) for any non-influential observation Z_k, the asymptotic distributions of $T_{\min,k,m}$ and $T_{\max,k,m}$ are the same as those in theorem 1, i.e. the distribution of the min- and max-statistic of a non-influential observation is not affected by the presence of influential observations. As such, a non-influential point can be identified as non-influential with high probability, i.e. the swamping effect can be overcome under the condition in result (a). Since $\tilde{F}_{\max,k}(\mathbb{Z}_n^{\inf}) \leqslant R^2_{\inf,n} d_{S_{\inf}}$ by expression (3.3), a sufficient condition for $\tilde{F}_{\max,k}(\mathbb{Z}_n^{\inf}) \to_p 0$ is that $R^2_{\inf,n} d_{S_{\inf}} \to_p 0$, which holds if $d_{S_{\inf}} = O_p(1)$ and $\delta_{\inf,n} \to 0$. This condition might be

violated, however, if $\delta_{\inf,n}$ does not vanish or some influential observations have large values in terms of E_t. This condition implies that deleting points with large values in E_t is helpful to alleviate the swamping effect.

For an influential observation Z_k, the max-unmask condition in result (b) gives the requirement on its signal strength for it to be identified as influential. Since $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ decreases with respect to m, this condition becomes weaker and easier to be satisfied, and Z_k is easier to be detected. This provides an opportunity to identify the influential observations that are masked by others, as long as we can make $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ sufficiently small. In fact, as shown in lemma 1.2 in the on-line supplementary materials, $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ can be very small if m is sufficiently large. Therefore, $T_{\max,k,m}$ has the advantage in overcoming the masking effect if m is large.

We do not assume explicitly any model for the influential observations, which makes the method proposed quite flexible. When a specific model is assumed on the influential observations, the conditions in theorem 2 have more explicit expression. We illustrate this point by investigating the mean shift model

$$\mathbf{X}_i^{\inf} = \mathbf{X}_i, \qquad Y_i^{\inf} = Y_i + c_i, \qquad i \in S_{\inf}, \tag{3.4}$$

where $c_i \neq 0$ and $(\mathbf{X}_i, Y_i)$s are independent and identically distributed non-influential observations from model (2.1), i.e. the influential observations are generated by contaminating the response of non-influential observations. To simplify the argument, we assume that $\mu_{xj} = \mu_y = 0$ and $\sigma_{xj} = 1$, $1 \leqslant j \leqslant p$, and consider the mean and variance of E_i. Specifically, $E_i = p^{-1}\|\dot{Y}_i \dot{\mathbf{X}}_i\|^2 = \dot{Y}_i^2 K_{p,ii}$, where $K_{p,ii} = p^{-1}\|\dot{\mathbf{X}}_i\|$. Because $E(K_{p,ii} - 1)^2 = O(p^{-2} l_p)$ in the proof of lemma 1, it follows that $E\{\max_{i \in S_{\inf}}(K_{p,ii} - 1)^2\} = O_p(n_{\inf} p^{-2} l_p)$ and consequently that $E(\max_{i \in S_{\inf}} K_{p,ii}) \leqslant 1 + O_p(n_{\inf}^{1/2} p^{-1} l_p^{1/2})$. By assuming that $n_{\inf}^{1/2} p^{-1} l_p^{1/2} = o(1)$, we see that $\max_{i \in S_{\inf}} K_{p,ii} = 1 + o_p(1)$. Suppose that $|c_i| \gg \log(n_{\inf})^{1/2}$ for $i \in S_{\inf}$. Thus, $E_i = (Y_i^{\inf})^2\{1 + o_p(1)\} = c_i^2\{1 + o_p(1)\}$, by noting that $Y_i \sim N(0, 1)$ and consequently $\max_{i \in S_{\inf}} |Y_i| = O_p\{\log(n_{\inf})^{1/2}\}$. Then, by expression (3.3), the condition in result (a) of theorem 2 becomes $(n_{\inf}/n)^2 \max_{i \in S_{\inf}} c_i^2 \to 0$. Since $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_{\mathrm{p}} 0$, as $m \to \infty$, by lemma 1.2 in the on-line supplementary materials, the max-unmask condition can be relaxed as $\min_{i \in S_{\inf}} |c_i| > \chi^2_{1-\alpha}(1)^{1/2} + \epsilon_0$ for sufficiently large m, which holds trivially as $|c_i| \gg \log(n_{\inf})^{1/2}$, $i \in S_{\inf}$.

We now formally formulate a multiple-testing problem to test the influentialness of individual observations with n null hypotheses $H_{0k}: Z_k$ is non-influential, $1 \leqslant k \leqslant n$. By result (b) of theorem 2 and the above discussion, we can estimate the set of the influential observations as

$$\hat{S}_{\max} = \{k : p_{\max,k,m} < q_k, 1 \leqslant k \leqslant n\},$$

where $p_{\max,k,m} = P\{\chi^2(1) > T_{\max,k,m}\}$ is the p-value under H_{0k} and the q_ks are determined by the specific procedure that is used to control the error rate. Here the q_ks can be independent of k, if we aim to control the familywise error rate by the Bonferroni test. Alternatively, the q_ks can depend on k, if we want to control the FDR at level α_0. For example, for the procedure in Benjamini and Hochberg (1995), q_k can be taken as the largest $p_{\max,(k)}$ such that $p_{\max,(k)} \leqslant k\alpha_0/n$, where $p_{\max,(1)} \leqslant p_{\max,(2)} \leqslant \ldots \leqslant p_{\max,(n)}$ are the ordered $p_{\max,k,m}$s. We now state the theory of using the Benjamini–Hochberg procedure and will use it later for numerical illustration, although other procedures developed for controlling the FDR can also be used.

Proposition 1. Suppose that the Benjamini–Hochberg procedure is used to control the FDR at level α_0. Assume that assumption 1 and the conditions in lemma 1 hold. Suppose that the max-unmask condition in result (b) of theorem 2 holds simultaneously for all $k \in S_{\inf}$ with $\alpha < \delta_{\inf,n}\alpha_0$.

Specifically, for any $m = m(n)$, if $P\{\min_{k\in S_{\inf}} E_k^{1/2} > \chi^2_{1-\alpha}(1)^{1/2} + \max_{k\in S_{\inf}} F^{1/2}_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) + \epsilon_0\} \to 1$, as $\min\{n, p\} \to \infty$, where $\epsilon_0 > 0$ is sufficiently small, then we have $P(\hat{S}_{\max} \supseteq S_{\inf}) \to 1$.

As discussed in remark 1 in the on-line supplementary materials, when n and m become large, we have $\max_k F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \leqslant a_0$ with probability tending to 1 for some small a_0. Proposition 1 shows that all the influential points will be identified as influential with high probability, i.e. the true positive rate (TPR) is well controlled. In addition, if $R^2_{\inf,n} d_{S_{\inf}} \to_{\mathrm{p}} 0$, by expression (3.3) and result (a) in theorem 2, there will be no swamping effect and then the statistic $T_{\max,k,m}$ under H_{0k} follows a $\chi^2(1)$ distribution. Let $\mathrm{FPR}(\hat{S}_{\max}) = |\hat{S}_{\max} \cap S^{\mathrm{c}}_{\inf}|/|S^{\mathrm{c}}_{\inf}|$ be the estimated FPR. When the Benjamini–Hochberg procedure is applied and there is no swamping effect, $\mathrm{FPR}(\hat{S}_{\max})$ will be controlled. However, the condition $R^2_{\inf,n} d_{S_{\inf}} \to_{\mathrm{p}} 0$ may fail if $\delta_{\inf,n}$ does not converge to 0. In this case, the FPR may be out of control.

To summarize, the detection procedure based on the max-statistic $T_{\max,k,m}$ is effective in overcoming the masking effect, but it is somewhat aggressive in that the FPR may not be controlled well without strong conditions. However, we point out that the procedure based on $T_{\max,k,m}$ is computationally efficient, compared with that based on $T_{\min,k,m}$ below.

3.2. Min-statistic $T_{\min,k,m}$ for the kth point

We have argued that the statistic $T_{\min,k,m}$ is effective in alleviating the swamping effect. We formally state this in the following theorem. Recall that $T_{\min,k,m}$ is a function of $(\mathbb{Z}_n, \mathbb{A}_m)$. It makes sense to investigate the behaviour of $T_{\min,k,m}$ given $\mathbb{A}_m$.

Theorem 3. Under assumption 1 and the assumptions of lemma 1, when there are influential observations, the following two conclusions hold.

(a) Suppose that Z_k is non-influential. For any $m = m(n)$ satisfying $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_{\mathrm{p}} 0$ when $\min\{n, p\} \to \infty$, it holds that $P(T_{\min,k,m} > t) \leqslant P\{\chi^2(1) > t\}$ for any $t \in \mathbb{R}$, as $\min\{n, p\} \to \infty$.

(b) Suppose that Z_k is influential. For any $\epsilon_0 > 0$ sufficiently small, define $\mathcal{I}_{\max,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) = \{E_k^{1/2} > \chi^2_{1-\alpha}(1)^{1/2} + F_{\max,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)^{1/2} + \epsilon_0\}$. As $\min\{n, p\} \to \infty$, it holds uniformly over $(m, \mathbb{A}_m)$ that

$$P\{T_{\min,k,m} > \chi^2_{1-\alpha}(1) | \mathbb{A}_m\} - P\{\mathcal{I}_{\max,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) | \mathbb{A}_m\} \geqslant 0.$$

Therefore, for any $m = m(n)$, if the following min-unmask condition holds,

$$P\{\mathcal{I}_{\max,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)\} \to 1, \qquad \text{as } \min\{n, p\} \to \infty,$$

then $P\{T_{\min,k,m} > \chi^2_{1-\alpha}(1)\} \to 1$, as $\min\{n, p\} \to \infty$.

By expression (3.3), $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_{\mathrm{p}} 0$ uniformly over $(m, \mathbb{O}_m)$, as $R^2_{\inf,n} d_{S_{\inf}} \to_{\mathrm{p}} 0$. In this case, conclusion (a) of theorem 3 can be strengthened as $T_{\min,k,m} \to_{\mathrm{d}} \chi^2(1)$ uniformly over $(m, \mathbb{A}_m)$, as $\min\{n, p\} \to \infty$. Moreover, by expression (3.3), it holds that $\mathcal{I}_{\max,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \supseteq \tilde{\mathcal{I}}_{\max,k}(\mathbb{Z}_n^{\inf}) := \{E_k^{1/2} > \chi^2_{1-\alpha}(1)^{1/2} + \tilde{F}_{\max,k}(\mathbb{Z}_n^{\inf})^{1/2} + \epsilon_0\}$. The min-unmask condition can be strengthened as $P\{\tilde{\mathcal{I}}_{\max,k}(\mathbb{Z}_n^{\inf})\} \to 1$, as $\min\{n, p\} \to \infty$, and conclusion (b) of theorem 3 can be strengthened as $P\{\min_{1\leqslant m\leqslant\infty} T_{\min,k,m} > \chi^2_{1-\alpha}(1)\} \to 1$, as $\min\{n, p\} \to \infty$.

When the influential observations are generated from the mean shift model (3.4), similarly to the discussion after theorem 2, we can see that the min-unmask condition holds with probability tending to 1 for any $1 \leqslant m \leqslant \infty$, if $|c_i| \gg \log(n_{\inf})^{1/2}$ as $i \in S_{\inf}$ and

$$|c_k| > \chi^2_{1-\alpha}(1)^{1/2} + \frac{n_{\inf}}{n} \max_{i\in S_{\inf}} |c_i| + \epsilon_0, \qquad k \in S_{\inf}.$$

Note that m plays a role in the conditions of theorem 3 and theorem 2. Compared with result (a) of theorem 2 where $\tilde{F}_{\max,k}(\mathbb{Z}_n^{\inf}) \to_p 0$ is required, condition (a) of theorem 3 is much weaker. As shown in lemma 1.2 in the on-line supplementary materials or the discussion just before Section 3.1, we see that $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_p 0$ when m is sufficiently large. Therefore, the statistic $T_{\min,k,m}$ is less sensitive to the swamping effect. In contrast, $F_{\max,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ is involved in the min-unmask condition (b), which is much stronger than the max-unmask condition (b) of theorem 2, i.e. an influential observation Z_k will not be identified as influential unless its signal is very strong. Thus, the min-statistic is efficient in preventing the swamping effect but may be conservative for identifying influential points. Combining with the result in Section 3.1 that the max-statistic $T_{\max,k,m}$ is effective in overcoming the masking effect but is aggressive, we conclude that the max-statistic $T_{\max,k,m}$ and the min-statistic $T_{\min,k,m}$ are complementary to each other.

If the min-unmask condition holds for all $k \in S_{\inf}$ simultaneously, then Z_k with $k \in S_{\inf}$ will be detected correctly, when a certain error control procedure is used. For example, similarly to proposition 1, with $\alpha = \delta_{\inf,n}\alpha_0$, one can show that the Benjamini–Hochberg procedure can correctly detect the influential observations. However, the min-unmask condition is very strong and may not be satisfied for all $k \in S_{\inf}$ simultaneously. We provide a sufficient condition for this condition to hold. Without loss of generality, assume that $S_{\inf} = \{1, \ldots, n_{\inf}\}$ and write $E_{(1)} \geqslant E_{(2)} \geqslant \ldots \geqslant E_{(n_{\inf})}$ ranking $E_i, 1 \leqslant i \leqslant n_{\inf}$, in decreasing order. Recall that $R_{\inf,n} = \delta_{\inf,n}/k_{\mathrm{sub}}$.

Proposition 2. Suppose that $E_{(n_{\inf})}^{1/2} > R_{\inf,n} E_{(1)}^{1/2} + \chi_{1-\alpha}^2(1)^{1/2} + \epsilon_0$ holds in probability tending to 1, as $\min\{n, p\} \to \infty$, where $\epsilon_0 > 0$ is sufficiently small. Then the min-unmask condition holds in probability simultaneously for all the influential points $k \in S_{\inf}$ and any $1 \leqslant m \leqslant \infty$, as $\min\{n, p\} \to \infty$.

The condition in proposition 2 is strong. When $\delta_{\inf,n} > 0$ and $E_{(1)}$ is large, proposition 2 requires that $E_{(n_{\inf})}$ is not too small but this condition may be violated easily. A remedy is to remove sequentially the influential observations that have been detected so far and then to apply the detecting procedure recursively on the remaining data, as we explain below.

To simplify the description, we introduce some notation. For any subset $U \subseteq \{1, \ldots, n\}$ with cardinality $n_U = |U|$ and any observation $Z_{k'}$ with $k' \in U$, we can draw at random with replacement subsets $A_{1,U}, \ldots, A_{m,U} \subset U \setminus \{k'\}$, with the same cardinality $n_{\mathrm{sub},U}$, where $n_{\mathrm{sub},U} < n_U$. Similarly to $T_{\min,k,m}$, we define $T_{\min,k',m}(U) = \min_{1 \leqslant r \leqslant m} n_{\mathrm{sub},U}^2 D_{r,k',U}$, where $D_{r,k',U} = p^{-1} \|\hat{\boldsymbol{\rho}}_{A_{r,U}^{(+k')}} - \hat{\boldsymbol{\rho}}_{A_{r,U}}\|^2$, $\mathbb{Z}_U = \{Z_i, i \in U\}$ and $\mathbb{A}_{m,U} = (A_{1,U}, \ldots, A_{m,U})$.

Denote by $B_{r,U}$ the indices of non-influential observations in $A_{r,U}$ and let $O_{r,U} = A_{r,U} \setminus B_{r,U}$, $1 \leqslant r \leqslant m$. Let $k_{\mathrm{sub},U}$ be such that $n_{\mathrm{sub},U} = n_U k_{\mathrm{sub},U} + 1$. Then, similarly to $F_{\min,k}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$, we define $F_{\min,k'}(\mathbb{Z}_U^{\inf}, \mathbb{O}_{m,U}) = \min_{1 \leqslant r \leqslant m} p^{-1} \|\Sigma_{t \in O_{r,U}} \hat{Y}_t \hat{\mathbf{X}}_t / (n_U k_{\mathrm{sub},U})\|^2$ with $\mathbb{Z}_U^{\inf} = \{Z_i, i \in S_{\inf} \cap U\}$ and $\mathbb{O}_{m,U} = (O_{1,U}, \ldots, O_{m,U})$, which denotes the minimum of the joint effect of influential observations with indices in U. Similarly we can define $F_{\max,k'}(\mathbb{Z}_U^{\inf}, \mathbb{O}_{m,U})$. Obviously, when $U = \{1, \ldots, n\}$, $T_{\min,k',m}(U)$, $F_{\min,k'}(\mathbb{Z}_U^{\inf}, \mathbb{O}_{m,U})$ and $F_{\max,k'}(\mathbb{Z}_U^{\inf}, \mathbb{O}_{m,U})$ are exactly the same as $T_{\min,k',m}$, $F_{\min,k'}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ and $F_{\max,k'}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m)$ respectively.

Generally, suppose that $E_{(i)}$s can be separated into several groups in successive order, i.e. $G_j = \{E_{(m_{j-1}+1)}, \ldots, E_{(m_j)}\}$, $j = 1, \ldots, \tau$, such that $0 = m_0 < m_1 < \ldots < m_\tau = n_{\inf}$. Denote $I_j = \{(m_{j-1}+1), \ldots, (m_j)\}$, $1 \leqslant j \leqslant \tau$. Let $M_0 = S_{\inf}$, $M_j = M_{j-1} \setminus I_j$ and $U_j = M_{j-1} \cup S_{\inf}^{\mathrm{c}}$, $1 \leqslant j \leqslant \tau$. For simplicity, we assume that the n_{sub,U_j}s are independent of j, denoted still as n_{sub}, and that the sufficient condition in proposition 2 holds for group G_js, i.e. the following inequality holds simultaneously in probability tending to 1, as $\min\{n, p\} \to \infty$,

$$E_{(m_j)}^{1/2} > R_{\inf,n} E_{(m_{j-1}+1)}^{1/2} + \chi_{1-\alpha}^2(1)^{1/2} + \epsilon_0, \qquad 1 \leqslant j \leqslant \tau, \tag{3.5}$$

which is referred to as the group-min-unmask condition for simplicity. Then, similarly to the argument of proposition 2, we see that the min-unmask condition holds simultaneously for any $Z_k, k \in I_j$, on the data set $\{Z_i, i \in U_j\}$, i.e. $P[\cap_{k\in I_j}\{E_k^{1/2} > F_{\max,k}^{1/2}(\mathbb{Z}_{U_j}^{\inf}, \mathbb{O}_{m,U_j}) + \chi^2_{1-\alpha}(1)^{1/2}\}] \to 1$ for any $1 \leqslant m \leqslant \infty$. Consequently $T_{\min,k,m}(U_j)$ with $k \in I_j$ will be larger than $\chi^2_{1-\alpha}(1)$ with high probability. If influential observations in $I_1, \ldots, I_{j-1}$ are detected correctly and removed sequentially, the influential observations in group I_j can be detected successfully with high probability. We remark that the group-unmask condition is much weaker than the condition in proposition 2.

This motivates us to consider the following multiround procedure. Define the set of influential observations identified in the jth round as

$$\hat{S}_{\min}^{j} = \{k : P\{\chi^2(1) > T_{\min,k,m}(\hat{U}_j)\} < q_k, Z_k \in \hat{U}_j\},$$

where q_k depends on the specific procedure that is used, similarly to the discussion in Section 3.1, $\hat{U}_j = \hat{U}_{j-1} \setminus \hat{S}_{\min}^{j-1}$ with $\hat{U}_0 = \{1, \ldots, n\}$ and $\hat{S}_{\min}^{0} = \emptyset$. Finally, we can estimate $S_{\inf}$ by $\hat{\mathbb{S}}_{\tau'} = \cup_{j=1}^{\tau'} \hat{S}_{\min}^{j}$ for some τ'. Let $\mathrm{FPR}(\hat{\mathbb{S}}_{\tau'})$ be the FPR that is associated with estimate $\hat{\mathbb{S}}_{\tau'}$.

Recall condition (a) of theorem 3, where $m = m(n)$ is such that $F_{k,\min}(\mathbb{Z}_n^{\inf}, \mathbb{O}_m) \to_{\mathrm{p}} 0$, as $\min\{n, p\} \to \infty$, with $k \in S_{\inf}^{\mathrm{c}}$. This condition ensures that the min-statistic for any $Z_k \in S_{\inf}^{\mathrm{c}}$ is smaller than $\chi^2(1)$ stochastically. When a multiround procedure is used, we need this condition to hold for each round. Specifically, we assume that $m = m(n)$ can be any series such that

$$\max_{\mathbb{Z}' \subseteq \mathbb{Z}_n^{\inf}} \max_{1 \leqslant k \leqslant n} F_{\min,k}(\mathbb{Z}', \mathbb{O}'_m) \to_{\mathrm{p}} 0, \tag{3.6}$$

as $\min\{n, p\} \to \infty$, where $\mathbb{O}'_m$ is defined similarly to $\mathbb{O}_m$. By expression (ii) of lemma 1.2 in the online supplementary materials, we see that result (3.6) holds as m is sufficiently large, regardless of the distribution of $\mathbb{Z}_n^{\inf}$. Then the FPR of the multiround procedure can be controlled as follows.

Proposition 3. Suppose that the FDR is controlled at level α_0 in each round, and that result (3.6) holds. Under assumption 1 and the conditions of lemma 1, for any fixed τ' such that $|\hat{\mathbb{S}}_{\tau'}^{c}| > c_0 n$ for some $c_0 > \delta_{\inf,n}/k_{\mathrm{sub}} + \delta_{\min}$, where $\delta_{\min} > 0$ is sufficiently small, it holds that $E\{\mathrm{FPR}(\hat{\mathbb{S}}_{\tau'})\} \leqslant \alpha_0$, as $\min\{n, p\} \to \infty$.

Although the above iterative procedure can improve the performance of the min-statistic for overcoming the masking effect, requiring only the weaker group-min-unmask condition in expression (3.5), the computation of this procedure will be more costly if the number of rounds τ' is large. However, larger τ' demands more intensive computing and may increase the FPR. If an early stopping strategy is adopted, it may still suffer from the masking effect.

As a quick summary, the test statistic $T_{\max,k,m}$ is more efficient in dealing with the masking effect, because the strength of the influential observations that is required by $T_{\max,k,m}$ in result (b) of theorem 2 is much weaker than the group-min-unmask condition (3.5) that is required by $T_{\min,k,m}$, when m is large. Moreover, any procedure based on $T_{\max,k,m}$ is computationally efficient, identifying the influential observations in just one round. However, $T_{\max,k,m}$ may suffer from the swamping effect if the strong condition (a) of theorem 2 is violated. In contrast, the estimate $\hat{\mathbb{S}}_{\tau'}$ based on the statistic $T_{\min,k,m}$ can maintain a good FPR at the expense of more intensive computation. Taking advantage of both statistics, we propose the following computationally efficient min-max-checking algorithm for identifying with high probability a clean set that contains no influential points and can serve as the benchmark for assessing the influence of other points.

3.3. Min-max-checking algorithm

We now present the following *min-max-checking algorithm* to combine the strength of the max- and min-statistic.

Step 1 (min-max-step): let $S_{\text{total}}^{(0)} = \{1, \ldots, n\}$ and fix $k_{\text{sub}} = \frac{1}{2}$. Repeat the following steps 1.1 and 1.2 until stopping for estimating a clean set.

1.1 (min-step): for the data indices in $S_{\text{total}}^{(t-1)}$, compute $\hat{M} = \{k : P\{\chi^2(1) > T_{\min,k}\} < \alpha_k, 1 \leqslant k \leqslant n\}$. Alternatively we may simply take $\hat{M}$ as the set of indices with the first l_0 smallest p-value for some small number l_0. Update $S_{\text{total}}^{(t)} \leftarrow S_{\text{total}}^{(t-1)} \setminus \hat{M}$.

1.2 (max-step): estimate $\hat{S}_{\max}$ as in Section 3.1 based on observations in $S_{\text{total}}^{(t)}$ and denote its complement $\hat{S}_{\max}^{\text{c}}$ as an estimate of the clean set. If $|\hat{S}_{\max}^{\text{c}}| \geqslant k_{\text{sub}} n$, then stop; otherwise, let $t \leftarrow t + 1$ and go to the min-step.

Step 2 (checking step): denote the estimated clean subset as $\mathcal{S}_{\text{c}}$. Check for each $k \in \mathcal{S} = \{1, \ldots, n\} \setminus \mathcal{S}_{\text{c}}$ whether the kth observation is influential.

In the min-max-step, this algorithm identifies with high probability a clean data set containing no influential points with cardinality at least $n/2$ by successively removing potential influential points. Here α_k is specified by the procedure that controls the error rate and can be determined in the same way as q_k in Section 3.1. The main rationale is, as argued, that the max-statistic $T_{\max,k,m}$ is aggressive in declaring influential whereas the min-statistic $T_{\min,k,m}$ is conservative. We first run a min-step to eliminate those influential observations with strong strength to alleviate the swamping effect. Combined with the efficiency of $T_{\max,k,m}$ in overcoming the masking effect, it is highly possible to obtain a clean set with a large size in one iteration. If the clean set is not sufficiently large, we run the min-step again to remove further influential observations with strong strength.

In the checking step, we denote $\mathcal{S}_{\text{c}}$ as the final clean set obtained by the min-max-step. Then its supplement, written as $\mathcal{S} = \{1, \ldots, n\} \setminus \mathcal{S}_{\text{c}}$, is an estimate of the set which contains all potential influential observations. However, $\mathcal{S}$ may still contain non-influential observations as the procedure for obtaining a clean set aims only to find a subset of the non-influential points. A further step is to check whether any point in $\mathcal{S}$ is truly influential if necessary. This step, however, is easy since we have now a clean data set. We now outline the exact procedure. For any Z_i, $i \in \mathcal{S}$, consider the data with indices in $\mathcal{S}_{\text{c}}$ and $\mathcal{S}_{\text{c}}^{(i)} = \mathcal{S}_{\text{c}} \cup \{i\}$. We then compute statistic $\mathcal{D}_i$ as in Section 2 where $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\rho}}^{(i)}$ are computed on data set $\mathcal{S}_{\text{c}}$ and $\mathcal{S}_{\text{c}}^{(i)}$ respectively. Since $\mathcal{S}_{\text{c}}$ is a good estimate of the clean data, this leave-one-out approach may still be effective for testing multiple null hypotheses in the form of $H_{0i} : Z_i$ is non-influential, $i \in \mathcal{S}$. Those whose corresponding hypotheses are rejected are labelled as influential observations.

Numerically, when the Benjamini–Hochberg procedure is used to control the error rate, we find that the min-statistic is not only robust to the swamping effect but also powerful, eliminating most influential observations of moderate or large effects in one go. For example, it is seen from Fig. S.2 of the on-line supplementary materials that the TPR of the min-statistic is quite insensitive to m and is only slightly worse than that of the max-statistic whereas its FPR is much smaller. As a result, the swamping effect in the following max-step is very weak and this max-step can remove further influential observations of weak signal strength. Thus, in most of the cases that we have considered, one iteration is all that is needed for the min-max-step. The resulting estimated clean set $\mathcal{S}_{\text{c}}$ contains most of the non-influential observations, i.e. the difference between $\mathcal{S}_{\text{c}}$ and true clean set $S_{\text{inf}}^{\text{c}}$ is small. Since the non-influential observations are averaged in the HIM statistic, when the difference between $\mathcal{S}_{\text{c}}$ and $S_{\text{inf}}^{\text{c}}$ is small, the critical value determined by $\chi^2(1)$ still gives reasonable results in the checking step. Our numerical study in

Section 4 shows that, indeed, the resulting procedure can still control the FDR at the desired level.

In the numerical study, we find that the min-max-step with just one iteration already leads to good results, and consequently the estimates are often identical with and without the checking step. In what follows we provide a high level theoretical analysis of the algorithm when the min-max-step is iterated once, while leaving the details to the on-line supplementary materials.

Without loss of generality, assume that $S_{\inf} = \{1, \ldots, n_{\inf}\}$ and write $E_{(1)} \geqslant E_{(2)} \geqslant \ldots \geqslant E_{(n_{\inf})}$ ranking $E_i, 1 \leqslant i \leqslant n_{\inf}$, in a decreasing order. Suppose that the influential observations are separated into the two groups, $G_{\mathrm{st}} = \{(k) : 1 \leqslant k \leqslant \tilde{m}_1\}$ and $G_{wk} = \{(k) : \tilde{m}_1 + 1 \leqslant k \leqslant n_{\inf}\}$, where $n_{\inf} - \tilde{m}_1 < (k_{\mathrm{sub}}^2 - \delta_1')n$ for some $0 < \delta_1' < k_{\mathrm{sub}}^2$, which stand for the indices of the observations with large and small values of E_t respectively.

Under conditions that are similar to those in Section 3.2, the min-statistic can identify G_{st} successfully and, under conditions that are similar to those in Section 3.3, the following max-step, applied to the reduced data, can identify the remaining influential observations. The details of the conditions that are required are presented in conditions (D1) and (D2) in the on-line supplementary materials.

Denote by $\hat{S}_{\min}$ and $\hat{S}_{\max}^{\mathrm{ng}}$ the indices of the observations that are labelled as influential by the min-step and the following max-step respectively. Specifically, $\hat{S}_{\min} = \{k : p_{\min,k} < q_k, 1 \leqslant k \leqslant n\}$, where $p_{\min,k} = P\{\chi^2(1) > T_{\min,k,m}\}$ is the p-value that is computed on the basis of the full data. And $\hat{S}_{\max}^{\mathrm{ng}} = \{k : p_{\max,k} < q_k, k \in \hat{S}_{\min}^{\mathrm{c}}\}$, where $p_{\max,k} = P\{\chi^2(1) > T_{\max,k,m}(\hat{S}_{\min}^{\mathrm{c}})\}$ is the p-value that is computed from the reduced data $\{Z_k : k \in \hat{S}_{\min}^{\mathrm{c}}\}$ and $T_{\max,k,m}(\hat{S}_{\min}^{\mathrm{c}})$ denotes the max-statistic applied to the reduced data. Write $\hat{S}_{\mathrm{mm}} = \hat{S}_{\min} \cup \hat{S}_{\max}^{\mathrm{ng}}$ as the estimated influential observations in the min- and max-step together.

Theorem 4. Consider the min-max-step in one iteration with the Benjamini–Hochberg procedure that is used to control the FDR at level α_0. Assume that assumptions 1–5 and conditions (D1) and (D2) in the on-line supplementary materials hold, where $m = m(n)$ is any series satisfying conditions (D1) and (D2). Then it follows that $P(\hat{S}_{mm} \supseteq S_{\inf}) \to 1$ and that $E\{\mathrm{FPR}(\hat{S}_{mm})\} \leqslant \alpha_0$, as $\min\{n, p\} \to \infty$.

The benefit of combining the min- and max-step can be seen by comparing conditions (D1) and (D2) with those required by the max- or min-step alone. For the min-step alone to detect all influential observations successfully, we need the min-unmask condition in proposition 2, i.e. $E_{(n_{\inf})}^{1/2} > R_{\inf,n} E_{(1)}^{1/2} + \chi_{1-\alpha}^2(1)^{1/2} + \epsilon_0$ holds with probability tending to 1, whereas, for the min-max-step, we need only group-min-unmask condition (3.5), which is weaker. Moreover, for the max-step alone to overcome the swamping effect, we need the condition $\tilde{F}_{\max,k}(\mathbb{Z}_n^{\inf}) \to_{\mathrm{p}} 0$ as $\min\{n, p\} \to \infty$, which holds when $R_{\inf,n}^2 d_{S_{\inf}} \to_{\mathrm{p}} 0$ as $\min\{n, p\} \to \infty$ according to expression (3.3), whereas, for the min-max-step, we need only a weaker condition $(R_{\inf,\tilde{m}_1}^{\mathrm{wk}})^2 E_{(\tilde{m}_1 - 1)} \to_{\mathrm{p}} 0$, where $E_{(\tilde{m}_1 - 1)} \leqslant E_{(1)} = d_{S_{\inf}}$ and $R_{\inf,\tilde{m}_1}^{\mathrm{wk}}$ defined in the on-line supplementary materials is smaller than $R_{\inf,n}$.

4. Simulation

We conduct extensive simulation to assess the performance of MIP. Towards this, we generate n observations from the linear model (2.1). The influential points are generated by replacing the first 10 points by points generated differently so that the resulting data set contains 10 influential points. In particular, we consider three examples. The first is constructed such that there is a strong masking effect whereas the second is generated to have a strong swamping effect. The magnitude of these effects is determined by a parameter μ, and the true β in these two examples

is sparse. For either example, we examine different combinations of n and p. We also consider a third example taken from Maronna (2011) in which β is random and non-sparse. For brevity, we summarize the main findings of the simulation study while leaving the details of the simulation to the on-line supplementary materials.

First, we set $(n, p) = (100, 1000)$ in examples 1 and 2 to assess the performance of MIP. In particular, we evaluate $\mathrm{TPR}_{\mathrm{inf}}$ as the TPR for influential observation detection, $\mathrm{FPR}_{\mathrm{inf}}$ as the FPR for detection, $\mathrm{ERR} = \|\hat{\beta} - \beta\|$ to measure the accuracy of the β estimate after influential points declared by MIP have been removed, $\mathrm{TPR}_{\mathrm{vs}}$ as the TPR for estimating the support of β, and $\mathrm{FPR}_{\mathrm{vs}}$ as the FPR for estimating the support of β. With some abuse of notation, the lasso estimate after influential points declared by MIP are removed is abbreviated as MIP. For comparison, we provide the corresponding quantities when HIM is used for outlier detection or when the lasso is fitted to the full data. The results are summarized in Table S.1 in the on-line supplementary materials and Fig. 2.

We then compare MIP with a few more competitors for the three examples. This is done by generating data sets for examples 1 and 2 with p equals $n/2$, n or $2n$ and n equals 100, 300 or 500 to assess how effective MIP is for different sample size dimensionality combinations. In particular, for influential point detection, we compare MIP with Cook's distance, DFFITS and the iterative procedure for outlier detection IPOD in She and Owen (2011) when $p < n$, and we compare MIP with HIM for $p > n$. For parameter estimation and identification of the support of β in examples 1 and 2, the performance is compared with penalized least absolute deviation (LAD) (Wang *et al.*, 2007) and a robust estimate named MM-Lasso (Smucler and Yohai, 2017). For example 3, we compare MIP with LAD and a robust estimator called S-Ridge (Maronna, 2011), as in this case the true β is non-sparse. The results are summarized in section 2 of the on-line supplementary materials.

4.1. Tuning parameters of the algorithm

Before presenting the results, first we briefly discuss how to choose k_{sub}, the relative size of the random subsets, and m, the number of the random subsets to implement the algorithm.

For k_{sub}, we recommend specifying it as an upper bound of the proportion of influential points in the data set. A reasonable choice is $\frac{1}{2}$ as we expect that, for any influential point identification method to work, the number of non-influential points should be larger than that of influential points. Additional simulation using $k_{\mathrm{sub}} = \frac{1}{3}$ or $k_{\mathrm{sub}} = \frac{2}{3}$ suggests that the results are quite insensitive to the choice of k_{sub} (see Fig. S.1 in the on-line supplementary materials).

Intuitively, the effects of the number of random subsets m for computing $T_{\max}$ and $T_{\min}$ are opposite. For $T_{\max}$, larger m leads to higher TPR and FPR, whereas, for $T_{\min}$, larger m produces results with lower TPR and FPR. The min-max-checking algorithm of MIP somehow combines their advantages, giving results with higher TPR and better control of FPR as m increases. This is confirmed numerically in Fig. S.2 of the supplementary materials where seven values of m ($m = 1, 5, 10, 50, 100, 500, 1000$) are investigated for its influence on using $T_{\max}$, $T_{\min}$ and MIP for influence diagnosis. It is seen that MIP is quite insensitive to the choice of m, especially so when $m \geqslant 50$.

In practice, however, it may still be useful to have a data-driven procedure for specifying m and we here present one. Our starting point is that, as m increases, the estimated set of influential observations becomes stable and so does the sum of $|\log(p\text{-value})|$ of all rejected hypotheses in the min-max- and checking step of the algorithm. We can plot this sum, which is denoted as $g(m)$, against m and identify a point, which is denoted as M_0, such that $g(m)$ becomes flat after M_0. In other words, M_0 is the elbow or knee point of the function $g(x)$. There are many

algorithms for finding M_0 and we have implemented a method that was suggested in Satopaa *et al.* (2011). The effect of this data-driven procedure for choosing m is illustrated in Fig. S.3 in the supplementary materials. Since we have found that the identified M_0 performs similarly to using $m = 100$ and $m = 1000$, all the simulations below have been conducted by using $m = 100$.

4.2. Summary of the simulation study

In terms of identifying influential points, across all simulation settings in the three examples, we can see that MIP controls the FDR at the nominal level and often has more power than HIM, Cook's distance and DFFITS for the settings where their FPRs are controlled. Also, in all the cases that were considered, MIP performs uniformly better than IPOD. In terms of parameter estimation we see that MIP outperforms MM-Lasso and penalized LAD for examples 1 and 2, and S-Ridge and LAD for example 3, usually by a larger margin. In terms of identifying the support of β in examples 1 and 2, we again see that MIP almost always outperforms its competitors especially in obtaining the smallest FPR in terms of variable selection.

We now briefly discuss the time and space complexity of implementing MIP. To compute the min- or the max-statistic, we need to sample m subsets of the data. For each observation and each subset, we compute HIM, which has a computational complexity $O(np)$. Thus, the total computational complexity is of the order $O(n^2 pm)$ that scales linearly with m; see Fig. S.10 in the on-line supplementary materials for some simulation results. We note that the computational time can be substantially reduced because the computation of the min- and max-statistics can be parallelized by using multiple processors: one for each of the m subsets. In contrast, the space complexity is of the order $O(n + p + m)$.

5. Real data analysis

In this section, we apply MIP to a microarray data set in which p is large and compare it with HIM. We also apply MIP to two small data sets in which p is small and compare it with Cook's distance and DFFITS. We remark that our theory may not apply to the two small data sets as their dimensionality can be seen fixed. Nevertheless, it is interesting to see how MIP performs in this classical set-up.

5.1. High dimensional data

As an illustration, we apply MIP to detect influential points in the microarray data from Chiang *et al.* (2006) which were previously analysed by Zhao *et al.* (2013). For this data set, we focus on 120 12-week-old male offspring that were selected for tissue harvesting from the eyes and for microarray analysis. The data set contains over 31042 different probe sets. Following Huang *et al.* (2006), we take the probe gene TRIM32 as the response. This gene is interesting as it was found to cause Bardet–Biedl syndrome, which is a genetically heterogeneous disease of multiple-organ systems including the retina (Chiang *et al.*, 2006). One question of interest in this data analysis is to find genes whose expressions are correlated with that of gene TRIM32. We followed Huang *et al.* (2006) to exclude probes that were not expressed in the eye or that lacked sufficient variation and we select $p = 1500$ genes that are mostly correlated with the probe of TRIM32. Therefore, the analysis has $p = 1500$ predictors and a sample size $n = 120$. Before further analysis, all the probes are standardized to have mean 0 and standard deviation 1 (Huang *et al.*, 2006). Applying the lasso to the full data by using the default setting of the `glmnet` function in R (Friedman *et al.*, 2010), we identify 15 significant variables and the l_2-norm of the estimated coefficient vector equals 0.097.

Applying HIM and MIP to these data with the FDR level at $\alpha = 0.05$, HIM finds 15 influential observations, whereas MIP obtains seven influential observations. Interestingly, the set of influential points by MIP is a subset of that by HIM. In Fig. 3, we plot the influential observations that were found by MIP as full circles and the extra influential observations by HIM as crossed circles, where the y-axis denotes the logarithm of the p-values that were obtained by using HIM as in Fig. 3(a) or using MIP as in Fig. 3(b). To make the plot more comparable, the checking step in the min-max-checking algorithm is applied to all observations such that we can obtain a p-value for each observation. From Fig. 3(b), we can see that the crossed circles that are identified by HIM as influential do not seem to have very small p-values.

To make further comparison, we use OLS estimation on the important variables found via the lasso, after applying either HIM or MIP, to the non-influential point set that was identified by HIM. We compare their Bayesian information criterion BIC-score defined as $\mathrm{BIC} = n \log(\mathrm{RSS}/n) + k \log(n)$ where RSS is the residual sum of squares, $n = 105$ is the sample size after removing the 15 influential points that were identified by HIM, and k is the number of variables used. Obviously, a model with a smaller BIC is preferred. Note that $k = 9$ if HIM is used and $k = 6$ if MIP is applied. Because of the set-up, this comparison favours HIM in some sense. It is found that $\mathrm{BIC} = -567.34$ if HIM is applied for influential point detection and $\mathrm{BIC} = -578.94$ if MIP is applied. Thus, MIP is potentially more effective for finding a better model than HIM as its BIC-value is smaller.

For the real data, of course it is not known which observations are influential. To assess the performance of HIM and MIP further, we artificially add influential points to the data set and evaluate whether they can find these points afterwards. Specifically, we first remove the influential points that were detected by each method and add 10 additional observations to the remaining data. This scheme gives a total of 115 observations for assessing HIM and 123 observations for MIP. The 10 added influential observations are generated as $X_{iS} = 1.1x_S + Z_S$, $X_{iS^c} = x_{S^c}$, $Y_i = 1.1y + \epsilon, 1 \leqslant i \leqslant 10$, where $Z \sim N(0, 0.01 I_p)$, S is a random subset of $\{1, \ldots, p\}$ consisting of 10 distinctive indices, Z_S is a subvector of Z with indices in S, (x, y) is chosen randomly from a non-influential point set identified by HIM and $\epsilon \sim N(0, 0.01)$ is independent of Z.

We apply MIP and HIM to the contaminated data defined above with the nominal FPR set as 0.05 in the Benjamni–Hochberg procedure and repeat the process 100 times. Then we compute the TPR and FPR of the two methods for identifying these artificial influential points. It turns out that MIP gives a TPR of 1 and an FPR of 0.008, whereas HIM gives a TPR of 1 and an FPR as high as 0.585. Obviously, HIM suffers seriously from the swamping effect that is caused by the addition of new influential observations, wheres MIP does not seem to be affected by newly added observations.

5.2. Low dimensional data sets

We now apply MIP to two classical data sets with small p that are used extensively in the literature as benchmark cases for influential diagnosis.

For the first data set, we examine the stack loss data in Brownlee (1965) consisting of $n = 21$ observations with $p = 3$. The covariates $X_i \in \mathbb{R}^3$ are air flow, cooling water inlet temperature and acid concentration, and the response Y_i is the stack loss. Several references identified five observations including cases 1, 2, 3, 4 and 21 as potential outliers (Rousseeuw and Leroy, 1987; Billor *et al.*, 2000; Nurunnabi *et al.*, 2014), usually by carefully examining the effect of deleting groups of observations on the OLS estimates. When applying MIP to this data set with the number of subsets m specified either as 20, 50 or 100, we always identify cases 1, 2 and 3 as outliers. If one believes that cases 1, 2, 3, 4 and 21 are the true outliers in some sense, then we

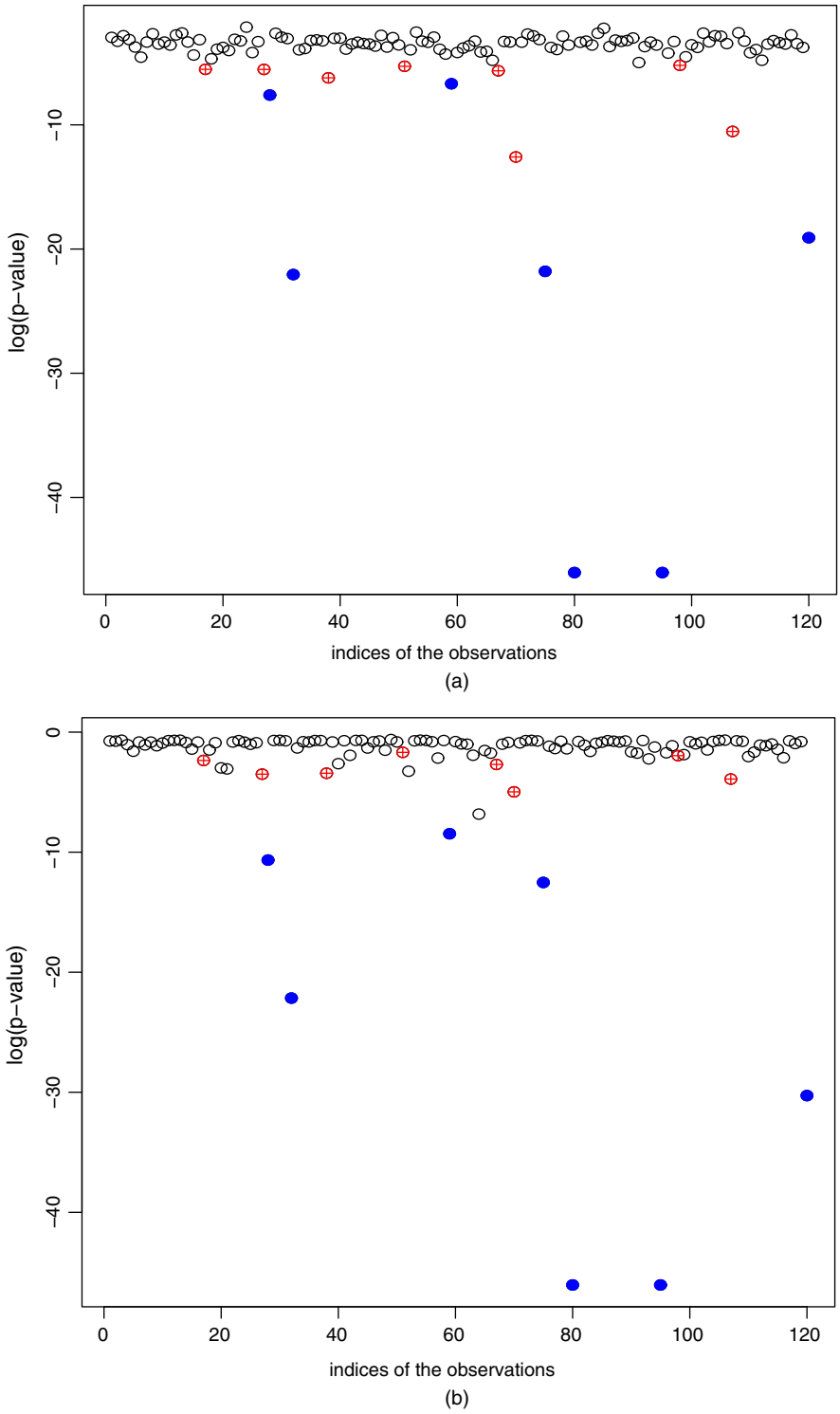

Fig. 3. Comparison between (a) HIM and (b) MIP

can see that the TPR of MIP is 0.6, whereas its FPR is 0. However, if we examine just Cook's distance by using leave-one-out observation, the TPR becomes 0 and the FPR becomes 0.0625. If the DFFITS-statistic is used for identifying outliers, then the TPR is 0 and the FPR is also 0. Neither Cook's distance nor DFFITS has any power in detecting these outliers.

For the second case, we look at a data set with $p = 3$ that is designed to have masking and swamping effects (Hawkins *et al*., 1984; Nurunnabi *et al*., 2014). There are $n = 75$ observations in total, the first 10 of which are specifically perturbed to be influential. Interestingly, after applying MIP, we find the first 13 observations as influential, meaning that the TPR of MIP is 1 and its FPR is 0.046. If we apply Cook's distance only, the TPR becomes 0 and the FPR is 0.0615, whereas the TPR becomes 0 and the FPR becomes 0 if we apply the DFFITS-statistic for outlier detection. For this example, MIP is much more powerful with a controlled FDR.

We point out that, to use MIP, we require $\min\{n, p\} \to \infty$, though the rates of n and p going to ∞ can be arbitrarily slow. From the analysis of the two low dimensional data sets above, however, we can see that MIP continues to provide useful results and at least is more competitive than examining Cook's distance or the DFFITS-statistic naively.

6. Discussion

We have proposed a novel procedure named MIP for multiple influential point detection in high dimensional spaces. The MIP procedure is intuitive, theoretically justified and easy to implement. In particular, by combining the strengths of the max- and min-statistics, the MIP framework proposed can overcome the masking and swamping effects that are notorious in influence diagnosis, and it can identify multiple influential points with prespecified accuracy in terms of FDR control which is empirically verified by extensive simulation.

Both HIM and MIP are based on the idea of measuring the change in marginal correlations when one observation is removed. The primary consideration for using the marginal correlation is due to its ubiquity in statistical analysis and the possibility of deriving rigorous theoretical results, as we have shown. But it need not be the only quantity that defines influence. Towards this, it will be interesting to explore the use of other quantities to define influence. In this paper, we have confined our attention to linear regression. An interesting topic for future research is to extend the idea to other models such as the generalized linear model. A major challenge, however, is to define a tractable influence measure that is similar to HIM.

Finally, we hope that this paper brings to the attention of the statistics community the importance of influence diagnosis and how one might think about defining influence and devising automatic procedures for assessing influence, in a theoretically justified fashion. With the rapid advances in 'big data' analytics, we believe that the issue of influence diagnosis will only become more relevant and we hope that this paper can serve as a catalyst to stimulate more research in this area.

Acknowledgements

We thank three reviewers, the Associate Editor and Joint Editor for their helpful comments that have led to a much improved paper.

Zhao's research is supported by National Natural Science Foundation of China, grants 11871104 and 11471030, and the Fundamental Research Funds for the Central Universities. Leng's research is partially supported by a Turing Fellowship under the Engineering and Physical Sciences Research Council grant EP/N510129/1.

References

Aggarwal, C. C. and Yu, P. S. (2001) Outlier detection for high dimensional data. *ACM Sigmod Rec.*, **30**, 37–46.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Billor, N., Hadi, A. S. and Velleman, P. F. (2000) Bacon: blocked adaptive computationally efficient outlier nominators. *Computnl Statist. Data Anal.*, **34**, 279–298.

Brownlee, K. A. (1965) *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.

Chatterjee, S. and Hadi, A. S. (1986) Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.*, **1**, 415–416.

Chiang, A. P., Beck, J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K. Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M. and Sheffield, V. C. (2006) Homozygosity mapping with SNP arrays identifies trim32, an e3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (bbs11). *Proc. Natn. Acad. Sci. USA*, **103**, 6287–6292.

Cook, R. D. (1977) Detection of influential observation in linear regression. *Technometrics*, **19**, 15–18.

Draper, N. R. and Smith, H. (2014) *Applied Regression Analysis*, 3rd edn. New York: Wiley.

Fan, J., Fan, Y. and Barut, E. (2014) Adaptive robust variable selection. *Ann. Statist.*, **42**, 324–351.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.

Filzmoser, P., Maronna, R. A. and Werner, M. (2008) Outlier identification in high dimensions. *Computnl Statist. Data Anal.*, **52**, 1694–1711.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.

Hadi, A. S. and Simonoff, J. S. (1993) Procedures for the identification of multiple outliers in linear models. *J. Am. Statist. Ass.*, **88**, 1264–1272.

Hawkins, D. M., Dan, B. and Kass, G. V. (1984) Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, **26**, 197–208.

Huang, J., Ma, S. and Zhang, C. H. (2006) Adaptive lasso for sparse high-dimensional regression. *Statist. Sin.*, **18**, 1603–1618.

Huber, P. J. and Ronchetti, E. M. (2009) *Robust Statistics*, 2nd edn. New York: Springer.

Imon, A. H. M. R. (2005) Identifying multiple influential observations in linear regression. *J. Appl. Statist.*, **32**, 929–946.

Lawrance, A. J. (1995) Deletion influence and masking in regression. *J. R. Statist. Soc.* B, **57**, 181–189.

Maronna, R. A. (2011) Robust ridge regression for high-dimensional data. *Technometrics*, **53**, 44–53.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006) *Robust Statistics: Theory and Methods*. New York: Wiley.

Nurunnabi, A. A. M. (2011) A diagnostic measure for influential observations in linear regression. *Communs Statist. Theory Meth.*, **40**, 1169–1183.

Nurunnabi, A. A. M., Hadi, A. S. and Imon, A. H. M. R. (2014) Procedures for the identification of multiple influential observations in linear regression. *J. Appl. Statist.*, **41**, 1315–1331.

Pan, J., Fung, W. and Fang, K. (2000) Multiple outlier detection in multivariate data using projection pursuit techniques. *J. Statist. Plannng Inf.*, **83**, 153–167.

Ro, K., Zou, C., Wang, Z. and Yin, G. (2015) Outlier detection for high-dimensional data. *Biometrika*, **102**, 589–599.

Roberts, S., Martin, M. A. and Zheng, L. (2015) An adaptive, automatic multiple-case deletion technique for detecting influence in regression. *Technometrics*, **57**, 408–417.

Rousseeuw, P. and Hubert, M. (2011) Robust statistics for outlier detection. *Data Minng Knowl. Discov.*, **1**, 73–79.

Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P. J. and van Zomeren, B. C. (1990) Unmasking multivariate outliers and leverage points. *J. Am. Statist. Ass.*, **85**, 633–639.

Satopaa, V., Albrecht, J. R., Irwin, D. E. and Raghavan, B. (2011) Finding a kneedle in a haystack: detecting knee points in system behavior. In *Proc. Int. Conf. Distributed Computing Systems, Minneapolis*, pp. 166–171. New York: Institute of Electrical and Electronics Engineers.

She, Y. and Owen, A. B. (2011) Outlier detection using nonconvex penalized regression. *J. Am. Statist. Ass.*, **106**, 626–639.

Shieh, A. D. and Hung, Y. S. (2009) Detecting outlier samples in microarray data. *Statist. Appl. Genet. Molec. Biol.*, **8**, 1–24.

Smucler, E. and Yohai, V. J. (2017) Robust and sparse estimators for linear regression models. *Computnl Statist. Data Anal.*, **111**, 116–130.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Velleman, P. F. and Welsch, R. E. (1981) Efficient computing of regression diagnostics. *Am. Statistn*, **35**, 234–242.

Wang, H., Li, G. and Jiang, G. (2007) Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econ. Statist.*, **25**, 347–355.
Welsch, R. E. (1982) Influence functions and regression diagnostics. *Modern Data Analysis*. New York: Academic Press.
Welsch, R. E. and Kuh, E. (1977) Linear regression diagnostics. *Technical Report 923-77*. Sloan School of Management, Massachusetts Institute of Technology, Cambridge.
Zhao, J., Leng, C., Li, L. and Wang, H. (2013) High-dimensional influence measure. *Ann. Statist.*, **41**, 2639–2667.
Zhu, H., Ibrahim, J. G. and Cho, H. (2012) Perturbation and scaled Cook's distance. *Ann. Statist.*, **40**, 785–811.
Zhu, H., Ibrahim, J. G., Lee, S. and Zhang, H. (2007) Perturbation selection and influence measures in local influence analysis. *Ann. Statist.*, **35**, 2565–2588.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Multiple influential point detection in high-dimensional regression spaces: supplementary materials'.