Multiple Influential Point Detection in High-Dimensional Spaces

Junlong Zhao  Chao Liu  Lu Niu  Chenlei Leng*

February 7, 2017

Abstract

Influence diagnosis is an integrated component of data analysis, but is severely under-investigated in a high-dimensional setting. One of the key challenges, even in a fixed-dimensional setting, is how to deal with multiple influential points giving rise to the masking and swamping effects. This paper proposes a novel group deletion procedure referred to as MIP by studying two extreme statistics based on a marginal correlation based influence measure. Named the Min and Max statistics, they have complimentary properties in that the Max statistic is effective for overcoming the masking effect while 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 proposed influential point detection procedure is simple to implement, 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.

Running Title: Multiple Influential Point Detection.

1 Introduction

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

An important assumption in linear regression is usually that the observations are all generated from the same model. In many applications, however, the data 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 alter the estimation of the regression coefficient and sway the outcome of 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 Had (1986); Imon (2005); Zhu et al. (2007, 2012); Nurunnabi et al. (2014), among many others. For high-dimensional models, Zhao et al. (2013) found that influential observations could negatively impact many methods recently developed for dealing with high-dimensionality, such as Lasso for variable selection (Tibshirani, 1996) and SIS for variable screening (Fan and Lv, 2008).

As a result, influence diagnosis has been long recognized as a central problem and routinely recommended 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, an excellent book on robust regression by Huber (2011) when \( p \) is fixed. Wang et al. (2007) and Fan et al. (2014), among others, devised robust methods for variable selection when heavy tailed noises are present, but no attempt was made to to quantify the influence of individual points, which can often be the main question of interest in practice. For multivariate data containing only \( X_i \)'s, Aggarwal and Yu (2001) proposed to find outliers in a high-dimensional space via projection, while Ro et al.

*Zhao is Associate Professor, School of Statistics, Beijing Normal University, China. Liu and Niu are graduate students, School of Systems and Mathematics, Beihang University, China. Leng is Professor, Department of Statistics, University of Warwick, UK. Leng is also affiliated with the Alan Turing Institute. Corresponding author: Chenlei Leng (Email: C.Leng@warwick.ac.uk).
used a robust covariance matrix estimator for defining distance for detecting outliers. She and Owen (2011) is among the first to study outlier detection in regression. Focusing on the mean shift model for $p < n$ problems, they did not show any theoretical guarantee for outlier detection. It is also found that empirically She and Owen’s method is outperformed by the approach proposed in this paper (Section 4).

When $p$ is fixed, there are many measures proposed for quantifying the influence of each observation, noticeably, 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). These measures have now been implemented in most statistical software such as R and SAS. Since these measures are all based on the ordinary least squares (OLS) estimation, they are not applicable to high-dimensional data. On the other hand, despite its obvious importance, the problem of influence diagnosis in a high-dimensional setting has received little attention. This is mainly due to the difficulty in establishing a coherent theoretical framework, even in a fixed-dimension setting, and lack of easily implementable procedures. Zhao et al. (2013) appears to be the first work on high-dimensional influence diagnosis. They proposed a new high-dimensional influence measure named HIM based on marginal correlations and established its asymptotic properties. The asymptotic theory further permits the development of a multiple testing based procedure for detecting influential points. 

Similar to many fixed dimensional measures, HIM is based on the idea of leave-one-out. That is, to quantify the influence of an observation, one compares a predefined measure evaluated on the whole dataset and the measure 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 in order to detect multiple ones. On the other hand, detecting multiple influential observations is much more challenging, due to the notorious “masking” and “swamping” effects (Hadi, 1993). Specifically, masking occurs when an influential point is not detected as influential, while swamping occurs when a non-influential point is classified as influential. In the language of multiple testing, masking is the problem of getting false negatives and swamping is the problem of getting false positives. To handle the masking and swamping effects in fixed dimensional models, many group deletion methods have been proposed (Rousseeuw and Zomeren, 1990; Hadi and Simonoff, 1993; Imon, 2005; Pan et al., 2000; Nurunnabi et al., 2014, Roberts et al., 2015). Dealing with these effects for high-dimensional data, however, is much more challenging and is currently an open problem.

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 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, while 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 Min-Max algorithm for obtaining a clean subset of the data that contains no influential points with high probability. This clean set of data is then served as the benchmark for assessing the influence of other observations, which permits one to control the false discovery rate 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 of any defined influence measures. On the contrary, our proposed MIP procedure is the first theoretically justified method and for the more challenging high-dimensional setting.

Before we proceed, we highlight the usefulness of the Max and Min statistics via an analysis of the microarray data in Section 4.3. Figure 1 plots the logarithms of the $p$-values associated with the Max statistic in (a) and the Min statistic in (b) of the observations, respectively. With a prespecified false discovery rate of 0.05, using the Min statistic, we identify a set of 7 influential observations, represented as the blue points in plot (a) and (b). It is interesting that the MIP procedure combining the strengths of the two statistics identifies the same set of 7 influential points. On the other hand, using the Max statistic, 4 additional observations, represented as red triangles in plot (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. On the other hand, the fact that the Min statistic gives the same set of influential points as MIP in plot (b) implies that there may not exist any masking effect in this data. Further analysis in Section 4.3 shows that the reduced data, obtained by removing the influential observations identified by MIP, results in a sparser model with a better fit, when Lasso is applied for model fitting.

The main flow of this paper is organized as follows. In Section 2, we review the high-dimensional influence measure in Zhao et al. (2013). In Section 3, based on the idea of random group deletion or leave-many-out, we...
propose Max and Min statistics for assessing extremeness and establish their theoretical properties. The Max and Min statistics for a given point are the maximum and the minimum quantity, respectively, of the influence measures defined over randomly subsampled data. 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, Theorem 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 the Min-Max algorithm can suitably combine their strengths.

Simulation results and data analysis, showing the competitive performance of MIP in comparison to HIM and the method of She and Owen [2011], are presented in Section 4. In Section 5, we provide further discussions. All the proofs are relegated to the Appendix. An R package implementing MIP, freely available on [http://www.warwick.ac.uk/chenleileng/research/](http://www.warwick.ac.uk/chenleileng/research/) now, will be deposited onto CRAN.

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

## 2 HIM, Masking and Swamping

### 2.1 Review of HIM

We first review the high-dimensional influence measure (HIM) in [Zhao et al. (2013)](http://www.warwick.ac.uk/chenleileng/research/) when \( \min\{p,n\} \to \infty \). Assume that the non-influential observations are i.i.d. from the following model

\[
Y_i = X_i^\top \beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}
\]

where \( Y_i \in \mathbb{R} \) is the response variable, \( X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p \) is the associated \( p \)-dimensional predictor vector, \( \beta \in \mathbb{R}^p \) is the coefficient vector, and \( \varepsilon_i \in \mathbb{R} \) is a normally distributed random noise with \( \text{cov}(X_i, \varepsilon_i) = 0 \). Denote \( \mu_y = E(Y_i), \sigma_y = (\text{var}(Y_i))^{1/2} \) and \( \mu_x = (\mu_{x1}, \cdots, \mu_{xp})^\top = E(X_i), \sigma_x = (\text{var}(X_{ij}))^{1/2}, 1 \leq j \leq p \).

The idea of HIM is to define the influence of a point by measuring its contribution to the average marginal correlation between the response and the predictors. Specifically, define the marginal correlation between variable \( j \) and the response as \( \hat{\rho}_j = \text{corr}(X_{ij}, Y_i) \). Given the data, we can obtain its sample estimate as \( \hat{\rho}_j = \frac{\sum_{i=1}^n (X_{ij} - \bar{\mu}_x)(Y_i - \bar{\mu}_y)}{n \hat{\sigma}_x \hat{\sigma}_y}, \) for \( j = 1, \ldots, p, \) where \( \bar{\mu}_x, \bar{\mu}_y, \hat{\sigma}_x, \hat{\sigma}_y \) are the sample estimates.
of $\mu_x, \mu_y, \sigma_x$ and $\sigma_y$, respectively. The sample marginal correlation with the $k$th observation removed is similarly defined as $\hat{\rho}_j^{(k)}$ for $1 \leq k \leq n$. HIM then measures the influence of the $k$th observation by comparing the sample correlations with and without this observation, defined formally as

$$D_k = p^{-1} \sum_{j=1}^{p} \left( \hat{\rho}_j - \hat{\rho}_j^{(k)} \right)^2, \quad 1 \leq k \leq n.$$ 

Intuitively, the larger $D_k$ is, the more influential the corresponding observation is. When there is no influential point and $\min\{n, p\} \rightarrow \infty$, under mild conditions, it is proved that $n^2D_k \rightarrow \chi^2(1)$, where $\chi^2(1)$ is the chi-square distribution with one degree of freedom. Based on this result, 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.

We now discuss why marginal correlation is attractive for defining influence. Cook’s distance and other classical influence measures rely on OLS which is infeasible in a high-dimensional setting whenever $p > n$. Constrained versions of OLS such as Lasso might seem useful, but their properties are extremely difficult to establish if the leave-one-out scheme is to be employed for studying influence. Even with additional assumptions such as sparsity on $\beta$, it is unlikely that the difference between the estimates with all the data and all the data but one can be rigorously established. On the other hand, an immediate advantage of using marginal correlation is that, as an ubiquitous quantity in statistics, it is well defined and more importantly tractable under this setting one can be rigorously established. On the other hand, an immediate advantage of using marginal correlation is, in some sense, equivalent to identifying those that influence the regression coefficient can also be detected by HIM, as we explain now.

Consider a simple mixture model in which $(X, Y)$ comes either from $Y = X^\top \beta + \epsilon$ (Model 1) with probability $1 - \theta$ or $Y = X^\top \beta_{\text{inf}} + \epsilon_{\text{inf}}$ (Model 2) with probability $\theta$, where $\theta \in [0, 1/2]$ is presumably small. With this setup, apparently, the aim of influence identification is to detect the points in Model 2. For simplicity, assume that $X, X_{\text{inf}}, \epsilon$, and $\epsilon_{\text{inf}}$ all have mean zero. Define

$$\rho_\theta := E(XY) = (1 - \theta)E(XY) + \theta E(X_{\text{inf}}Y_{\text{inf}}) = (1 - \theta)\text{cov}(X)\beta + \theta \text{cov}(X_{\text{inf}})\beta_{\text{inf}},$$

which is a function of $\theta$ whenever $\text{cov}(X)\beta \neq \text{cov}(X_{\text{inf}})\beta_{\text{inf}}$. By deleting one observation from the data as in HIM or multiple observations as in the MIP method, the empirical estimate $\hat{\rho}_\theta$ of $\rho_\theta$ changes as $\theta$ changes. This change can be fully exploited to identify influential points. More specifically, when $\text{cov}(X) = \text{cov}(X_{\text{inf}}) = \Sigma$ but $\beta \neq \beta_{\text{inf}}$, we have $\rho_\theta = \Sigma \beta_\theta$ where $\beta_\theta = (1 - \theta)\beta + \theta \beta_{\text{inf}}$. There is a one-to-one mapping between $\rho_\theta$ and $\beta_\theta$. The change in marginal correlation $\rho_\theta$ indicates a change in $\beta_\theta$ after re-scaled by $\Sigma$. Finding observations that influence marginal correlation is, in some sense, equivalent to identifying those that influence the regression coefficient. Furthermore, when there are abnormal points from covariates in that $\text{cov}(X) \neq \text{cov}(X_{\text{inf}})$ but $\beta = \beta_{\text{inf}}$, we can write $\rho_\theta = \Sigma \theta \beta$ where $\Sigma \theta = (1 - \theta)\text{cov}(X) + \theta \text{cov}(X_{\text{inf}})$. Again, there is a one-to-one correspondence between $\rho_\theta$ and $\Sigma \theta$. Identifying points that are abnormal in $\rho_\theta$ is equivalent to finding points abnormal in the covariates. In summary, the marginal correlation based measures can find influential points in the response, in the covariates, and in the coefficient, and HIM can be viewed as a screening method in this sense.

### 2.2 The effect of masking and swamping

Since HIM is based on the leave-one-out idea, the derived $\chi^2(1)$ distribution is invalid whenever there are one or more influential points. That is, 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 usefulness of any leave-one-out procedures. To appreciate how masking and swamping effects negatively impact the performance of HIM, we quickly look at Example 1 and 2 in Section 4. The data are generated such that there exists 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 $\mu$. Figure 2 presents a comparison of HIM in Zhao et al. (2013) and the proposed MIP method proposed in this paper for detecting influence, when the nominal level used for declaring influential in the Benjamini-Hochberg procedure is set at $\alpha = 0.05$.

From plot (a) of Figure 2, we see that the true positive rates (TPRs) of HIM are much lower than those of MIP; that is, 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 $\mu$ becomes large; that is, HIM identifies much more non-influential points as influential, meaning that HIM also suffers from the swamping effect. From plot (b), we see that HIM suffers from the swamping effect greatly, as the FPRs can be very close to 1 for large $\mu$. On the other hand, for both examples, the FPRs of the MIP procedure are controlled well below the nominal level while its TPRs are monotone functions of $\mu$ and eventually become one for large $\mu$.

Figure 2: Performance comparison between HIM and MIP. TPR: True positive rate; FPR: False positive rate. The nominal FPR is set at $\alpha = 0.05$, corresponding to the horizontal dotted grey line.

(a) Masking effect example (Example 1)  
(b) Swamping effect example (Example 2)

3 A Random Group Deletion Procedure

As discussed before, 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. Group deletion has also 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 and a good estimate of $\beta$ is necessary. However, in the high dimensional setting considered in this paper, extending these methods is challenging.

For our random group deletion procedure, the subsets are chosen with replacement uniformly at random. Thus, the marginal correlations based on these subsets can be seen as some kind of perturbations to the marginal correlations based on the whole sample. Their extremeness is summarized by two extremal statistics whose theoretically properties can be studied. Existing group deletion procedures are not employed in a way similar to how we define our statistics which are theoretically tractable.

Recall that $S_{\text{inf}}$ and $S_{\text{cinf}}$ denote the indices of influential and non-influential observations such that $S_{\text{inf}} \cup S_{\text{cinf}} = \{1, \ldots, n\}$. Let $|S_{\text{inf}}| = n_{\text{inf}}$ be the size of influential point set and $|S_{\text{cinf}}| = n - n_{\text{inf}}$ be the number of non-influential points. Write $Z_k = (X_k, Y_k), 1 \leq k \leq n$ as the $k$th data point. For any fixed $k$, to check whether $Z_k$ is influential or not, we draw uniformly at random with replacement some subsets $A_1, \cdots, A_m \subset \{1, \cdots, n\}/\{k\}$; that is, these subsets do not include $Z_k$. The choice of $m$ will be discussed in Section 3.3 and Section 4. Write $|A_r| = n_{\text{sub}} - 1$ where $n_{\text{sub}} = k_{\text{sub}}n + 1$ for some $k_{\text{sub}} \in (0, 1)$. These subsets are repeatedly drawn in the hope that there exists some subset that contains no influential observations. If such a clean set can be found, then
the statistic associated with any non-influential point has the $\chi^2(1)$ distribution as HIM. A conservative choice for $k_{sub}$ is $1/2$, because the number of non-influential points is usually larger than that of the influential points. Formally, we make the following assumption on $n_{inf}$ and $k_{sub}$.

(C1) Denote $\delta_{inf,n} = n_{inf}/n$ which is allowed to vary with $n$. Assume $0 \leq \delta_{inf,n} < 1/2 - \delta_1$ for some $\delta_1 > 0$ independent of $n$. We take $k_{sub} > \limsup n_{inf,n} + \delta_1$.

Assumption (C1) allows $\lim \delta_{inf,n} \to 0$. For $1 \leq r \leq m$, let $B_r$ be the subset of non-influential observations in $A_r$ and denote its size as $N_{B_r} = |B_r|$. Under (C1), we have $\lim \min_{1 \leq r \leq m} N_{B_r} > \delta_1 n$, that is, for any subset $A_r$, the number of non-influential observations does not vanish.

For $1 \leq r \leq m$, 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 $r$th random subset $A_r$ as

$$D_{r,k} = p^{-1}\|\hat{\rho}_{A_r^{(k)}} - \hat{\rho}_{A_r}\|^2, \quad 1 \leq r \leq m,$$

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

$$T_{\min,k} = \min_{1 \leq r \leq m} n_{sub}^2 D_{r,k}, \quad T_{\max,k} = \max_{1 \leq r \leq m} n_{sub}^2 D_{r,k}.$$

We name them the Min and Max statistic respectively as they measure the extremeness of the influence measures based on randomly sample data. Note that the statistics defined here, using Euclidean norm, are invariant to the rotation of the covariates and to the scale translation of the response.

To establish the asymptotic behaviours of $T_{\min,k}$ and $T_{\max,k}$, we first study the behaviour of a key quantity $J_{\max,n} = \max_{1 \leq r \leq m} J_r$ in which $J_r$ is defined as

$$J_r = p^{-1} \sum_{j=1}^p \frac{1}{N_{B_r}} \sum_{t \in B_r} \hat{Y}_t X_{tj}^\top = p^{-1} \sum_{j=1}^p \hat{Y}_t \hat{X}_t^\top,$$

where $\hat{Y}_t = \hat{\sigma}_y^{-1}(Y_t - \hat{\mu}_y)$, $\hat{X}_t = \hat{D}_x^{-1}(X_t - \hat{\mu}_x)$, $1 \leq t \leq n$, and $\hat{D}_x$ is the estimate of $D_x = \text{diag}(\sigma_{x1}, \ldots, \sigma_{xp})$, a diagonal matrix in $\mathbb{R}^{p \times p}$. By definition, $J_r$ is the square of $\ell_2$ norm associated with the non-influential observations in $A_r$ only and is therefore unknown. Denote $\hat{X}_t = D_x^{-1}(X_t - \mu_x)$ as the population version of $\hat{X}_t$ and note that $\hat{Y}_t$ is the population version of $Y_t$. Without loss of generality, we assume in model (2.1) that $\mu_y = \mu_x = 0$ and $\sigma_y = \sigma_x = 1$, $1 \leq j \leq p$, respectively. Moreover, we make the following assumptions.

(C2) For $1 \leq j \leq p, 1 \leq s \leq q$, $\rho_{js}$ is constant and does not change as $p$ increases.

(C3) For the covariance matrix of the covariates $\Sigma = \text{cov}(X_i)$ with eigen-decomposition $\Sigma = \sum_{j=1}^p \lambda_j u_j u_j^\top$, we assume $l_p = \sum_{j=1}^p \lambda_j^2 = O(p^r)$ for some $0 \leq r < 2$.

(C4) The predictor $X_i$ follows a multivariate normal distribution and the random noise $\varepsilon_i$ follows a multivariate normal distribution with mean zero and an unknown variance.

(C5) Let $(Q_y, R_y) = ((\hat{\mu}_y - \mu_y)/\sigma_y, \sigma_y/\hat{\sigma}_y - 1)$, $S_{Qy} = \limsup_{n \to \infty} E(n^{1/2}Q_y)$ and $S_{Ry} = \limsup_{n \to \infty} E(n^{1/2}R_y)$. Assume that $S_{Qy}$ and $S_{Ry}$ are finite. Furthermore, there exist constants $0 < K, C < \infty$, independent of $n$ and $p$, such that for any $t > 0$,

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

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

Assumptions (C2)-(C4) are also made in Zhao et al. (2013). Since it is assumed that $\sigma_{xj} = 1, 1 \leq j \leq p$, we have $\text{tr}(\Sigma) = p$ and consequently it holds that $l_p \leq p^2$ by Cauchy-Schwarz inequality. When $l_p = p^2$, $\Sigma$ is a
degene rate matrix with rank one and (C3) rules out this case. On the other hand, (C3) applies when the largest
eigenva lue of $\Sigma$ is bounded. Assumption (C5) is similar to but stronger than (C.4) of Zhao et al. (2013), where
only eighth moments of $n^{1/2}(\hat{\mu}_y - \mu_y)$ and $n^{1/2}(\hat{\sigma}_y/\sigma_y - 1)$ are required. In Assumption (C5), $n^{1/2}(\hat{\mu}_x_j - \mu_x)$ is
assumed to have sub-Gaussian tails and $n^{1/2}(\hat{\sigma}_x_j/\sigma_x - 1)$’s have sub-exponential tails. This assumption is satisfied for the sample mean and the sample variance under the normality of $(X_i, Y_i)$’s. As alternatives to the sample estimates, robust estimates of $\mu_x, \mu_y$, $\sigma_x$, and $\sigma_y$ can also be used in practice. For example, we can estimate $\mu_x$ and $\mu_y$ by the sample median and $\sigma_x$ by the median absolute deviation (MAD) estimator, respectively. These estimates satisfy Assumption (C5) by noting the normality of $(X_i, Y_i)$’s. These robust estimates are the quantities used in our numerical examples.

We now quantify the magnitude of $J_{\max,n}$, the maximum effect of the non-influential points, which is a key
quantity for establishing the asymptotic properties of the Min and Max statistics.

**Lemma 1.** Assume that the non-influential observations satisfy (C2)-(C4) and that (C1) and (C5) hold. Assume further $\xi_{n,p} = n^{-1/2}(\log p)(\log n)(\log(np)) \to 0$. Then for any $1 \leq m \leq \infty$,

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

Obviously, $\xi_{n,p} \to 0$ if $n^{-1/4+\epsilon_0} \log p \to 0$ for some sufficiently small $\epsilon_0 > 0$. Here the number of the
subsamples $m$ is allowed to grow to $\infty$ to help us understand the approach as explained in the next section,
in which we always need $m$ to be large. Based on Lemma 1 we have the following conclusion when
there is no influential observation.

**Theorem 1.** Suppose that all observations are non-influential. Under the assumptions of Lemma 1 it holds that, for any $1 \leq k \leq n$, $T_{\min,k} \to_d \chi^2(1)$ and $T_{\max,k} \to_d \chi^2(1)$.

Theorem 1 seems surprising at first glance, since we always have $T_{\min,k} \leq T_{\max,k}$. An explanation is in
place. It will be shown that $D_{r,k}$ can be decomposed into two parts. The first part, depending on the quantity $E_k$ defined in the next paragraph, 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 $T_{\min,k}$ and $T_{\max,k}$ 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 $T_{\max,k}$ or $T_{\min,k}$ is larger than $\chi^2(1)$, the $(1-\alpha)100\%$ quantile of the
$\chi^2(1)$ distribution, for some prespecified $\alpha$ such as 0.05, we declare that there exist outliers.

Recall that $B_r$ is the set consisting of 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 $1 \leq r \leq m$, it is obvious that $O_r \subseteq S_{\inf} \setminus \{k\}$, the latter equal to $S_{\inf}$ if
$k \in S_{\inf}^c$. Since $|A_r| = n_{\text{sub}} - 1 = k_{\text{sub}} n$, similar to the proof of Theorem 1 we have

$$n^2_{\text{sub}} D_{r,k} = p^{-1} \| \hat{\rho} - \hat{\rho}^{(k)} \|^2 = p^{-1} \left\| \frac{1}{n_{\text{sub}} - 1} \sum_{t \neq k, t \in A_r} \hat{Y}_t \hat{X}_t^\top - \hat{Y}_k \hat{X}_k^\top \right\|^2$$

$$= p^{-1} \left\| \frac{1}{nk_{\text{sub}}} \sum_{t \in B_r} \hat{Y}_t \hat{X}_t^\top + \frac{1}{nk_{\text{sub}}} \sum_{t \in O_r} \hat{Y}_t \hat{X}_t^\top - \hat{Y}_k \hat{X}_k^\top \right\|^2$$

$$= p^{-1} \| W_{\text{inf},k,r} + W_{\text{non},k,r} - \hat{Y}_k \hat{X}_k^\top \|^2, \quad (3.1)$$

where $W_{\text{inf},k,r} = \sum_{t \in O_r} \hat{Y}_t \hat{X}_t^\top / nk_{\text{sub}}$ and $W_{\text{non},k,r} = \sum_{t \in B_r} \hat{Y}_t \hat{X}_t^\top / nk_{\text{sub}}$ are associated with influential and non-
influential observations, respectively. Define

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

which represents the effect of the $k$-th observation $Z_k$. Let

$$F_{\min,k} = \min_{1 \leq r \leq m} p^{-1} \| W_{\text{inf},k,r} \|^2 \quad \text{and} \quad F_{\max,k} = \max_{1 \leq r \leq m} p^{-1} \| W_{\text{inf},k,r} \|^2$$

quantify the maximum and minimum joint effect of the influential observations, respectively. The asymptotic behavior of $T_{\max,k}$ and $T_{\min,k}$ depends on the magnitude of $E_k$, $F_{\min,k}$ and $F_{\max,k}$ when multiple influential
observations are present. See Theorem 2 in Section 3.1 and Theorem 3 in Section 3.2. We state the properties
of $T_{\max,k}$ and $T_{\min,k}$ separately.
3.1 Max statistic $T_{\text{max},k}$ for the $k$th point

In Theorem 1 we derive the null distribution of $T_{\text{max},k}$ and $T_{\text{min},k}$ when there is no influential point. We now study $T_{\text{max},k}$ when there are influential observations and develop the corresponding detection procedure. Recall $n_{\text{inf}} = n_{\text{inf},n}$ and $k_{\text{sub}} > 0$ in (C1). Denote $\delta_{\text{inf},n}/k_{\text{sub}} = R_{\text{inf}}$, the ratio of $|S_{\text{inf}}|$ over $|A_r|$, and let $d_S = \max_{t \in S} E_t$, for any $S \subseteq S_{\text{inf}}$. Simple calculation in the proof of Theorem 2 shows $F_{\text{max},k} \leq \delta_{\text{inf},n}^2 d_{\text{sub}}$. We have the following results for $T_{\text{max},k}$.

**Theorem 2.** Under the assumptions of Lemma 1 when there are influential observations, the following two conclusions hold.

(i) Suppose further $F_{\text{max},k} \rightarrow 0$. If observation $k$ is non-influential, that is, $k \in S_{\text{inf}}^c$, then both $T_{\text{min},k}$ and $T_{\text{max},k}$ converge to $\chi^2(1)$ in distribution.

(ii) For an influential point $k \in S_{\text{inf}}$, if

$$
\text{Max-Unmask Condition : } E_{k}^{1/2} > \left(\chi_{1-\alpha}^2(1)\right)^{1/2} + F_{\text{min},k}^{1/2}
$$

holds for some small prespecified $\alpha > 0$ where $\chi_{1-\alpha}^2(1)$ is the $100(1-\alpha)$% quantile of a $\chi^2(1)$ distribution, then $P(T_{\text{max},k} > \chi_{1-\alpha}^2(1)) \rightarrow 1$. In addition, it holds that $F_{\text{min},k} < a_0^2 < \infty$ for some $a_0 > 0$.

Under the condition in (i), for any non-influential observation $Z_k$, the asymptotic distributions of $T_{\text{min},k}$ and $T_{\text{max},k}$ are the same as those in Theorem 1. That is, the distributions of the Min and Max statistics of a non-influential observation are not affected by the presence of influential observations. As such, a non-influential point can be identified as non-influential with high probability. That is, the swamping effect can be overcome under the condition in (i). Since $F_{\text{max},k} \leq \delta_{\text{inf},n}^2 d_{\text{sub}}$, a sufficient condition for $F_{\text{max},k} \rightarrow 0$ is that $\delta_{\text{inf},n}^2 d_{\text{sub}} \rightarrow 0$, which holds if $d_{\text{sub}} < C < \infty$ and $\delta_{\text{inf},n} \rightarrow 0$. This condition might be violated, however, if $\delta_{\text{inf},n}$ 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 (ii) gives the requirement on its signal strength for it to be identified as influential. As $a_0$ decreases, the condition becomes weaker and easier to be satisfied, and $Z_k$ is easier to be detected. This provides opportunity to identify the influential observations that are masked by others, as long as we can make $a_0$ small enough. In fact, as argued below, $a_0$ can be very small if $m$ is sufficiently large.

Now, we discuss the upper bound $a_0$ in (ii) of Theorem 2. Recall that $O_r$ denotes the indices of the influential observations in $A_r$ and note $|O_r| \leq n_{\text{inf}}$. Then we have

$$
F_{\text{min},k} = p^{-1} \min_{1 \leq r \leq m} ||W_{\text{inf},k,r}||^2 \leq \min_{1 \leq r \leq m} \left[ \left( \frac{|O_r|}{nk_{\text{sub}}} \right)^2 \max_{t \in S_{\text{inf}}} E_t \right].
$$

Define $N_{O,m} = \min_{1 \leq r \leq m} |O_r|$. By allowing $m = \infty$, it is easy to see that $N_{O,m}$ is a decreasing function of $m$ with $\lim_{m \rightarrow \infty} N_{O,m} = 0$, since there are many subsets $A_r$ that contain no influential observations under assumption (C1), i.e. $|O_r| = 0$. Therefore, $\lim_{m \rightarrow \infty} F_{\text{min},k} = 0$. Of course, in practice $m = \infty$ is not achievable. Assume further $d_{\text{sub}} = \max_{t \in S_{\text{inf}}} E_t < C < \infty$. Then $F_{\text{min},k} \leq C(N_{O,m}/(nk_{\text{sub}}))^2$, which will be small for large $m$ and $n$. If $d_{\text{sub}}$ is unbounded but $d_{\text{sub}}/(nk_{\text{sub}}) < C < \infty$ for some $0 < \delta < 1$, we have $F_{\text{min},k} \leq C N_{O,m}^2/(nk_{\text{sub}})^2$, which converges to 0, as $m, n \rightarrow \infty$. Generally, when $m$ and $n$ are large, $a_0$ will be small under some mild conditions.

Therefore, $T_{\text{max},k}$ has advantages in overcoming the masking effect if $m$ is large.

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

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

where $p_{\text{max},k} = P(\chi^2(1) > T_{\text{max},k})$ is the $p$-value under $H_{0k}$ and $q_k$‘s are determined by the specific procedure used to control the error rate. Here $q_k$‘s can be independent of $k$, if we aim to control the familywise error rate by the Bonferroni test. Alternatively, $q_k$‘s can depend on $k$, if we want to control the false discovery rate (FDR) at level $\alpha_0$. For example, for the procedure in [Benjamini and Hochberg 1995], $q_k$ can be taken as the largest $p_{\text{max},(k)}$ such that $p_{\text{max},(k)} \leq k\alpha_0/n$, where $p_{\text{max},(1)} \leq p_{\text{max},(2)} \leq \cdots \leq p_{\text{max},(n)}$ are the ordered $p_{\text{max},k}$‘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 FDR can also be used.

**Proposition 1.** Suppose that the Benjamini-Hochberg procedure is used to control FDR at level $\alpha_0$. If the Max-Unmasking condition in (ii) of Theorem 2 holds with $\alpha < \delta_{\inf,n,\alpha_0}$ but with $F_{\min,k}$ replaced by the constant $a_0^2$ defined there, then under the conditions in Lemma 1, we have $P(S_{\max} \geq S_{\inf}) \to 1$.

Note that $a_0$ discussed further after Theorem 2 is independent of $k$. Proposition 1 shows that all the influential points will be identified as influential with high probability. That is, the true positive rate is well controlled. In addition, if $\delta_{\inf,n,d_{\inf}} \to 0$, by (i) in Theorem 2 there will be no swamping effect and then the statistic $T_{\max,k}$ under $H_{\inf}$ follows $\chi^2(1)$ distribution. Let $FPR(\hat{S}_{\max}) = |\hat{S}_{\max} \cap S_{\inf}|/|S_{\inf}|$ be the estimated FPR. When the Benjamini-Hochberg procedure is applied and there is no swamping effect, $FPR(\hat{S}_{\max})$ will be controlled. However, the condition $\delta_{\inf,n,d_{\inf}} \to 0$ is strong and it may fail if $\delta_{\inf,n}$ does not converge to zero. In this case, FPR may be out of control.

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

### 3.2 Min statistic $T_{\min,k}$ for the $k$th point

We have argued that the statistic $T_{\min,k}$ is effective in alleviating the swamping effect. We formally state this in the following theorem.

**Theorem 3.** Under the assumptions of Lemma 1, the following two conclusions hold.

(i) Assume $F_{\min,k} \to 0$. For any non-influential point $k \in S_{\inf}^c$, it holds that $T_{\min,k} \to d \chi^2(1)$.

(ii) For any influential $Z_k$, if

\[
\text{Min-Unmask Condition: } E_{k}^{1/2} > E_{\max,k}^{1/2} + (\chi^2_{\alpha}(1))/2 \]

holds, then $P(T_{\min,k} > \chi^2_{\alpha}(1)) \to 1$, where $\alpha > 0$ is a small constant.

Compared with (i) of Theorem 2 where $F_{\max,k} \to 0$ is required, the condition in (i) of Theorem 3 is much weaker. As discussed in Section 3.1, $F_{\min,k} \to 0$ when $\min\{m, n\} \to \infty$. Therefore, the statistic $T_{\min,k}$ is less sensitive to the swamping effect. On the other hand, $F_{\max,k}$ is involved in the Min-Unmask Condition in (ii), which is much stronger than the Max-Unmask Condition in (ii) of Theorem 2. That is, 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 $T_{\max,k}$ is effective in overcoming the masking effect but is aggressive, we conclude that the Max statistic $T_{\max,k}$ and the Min statistic $T_{\min,k}$ 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 certain error control procedure is used. For example, similar 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, we assume $S_{\inf} = \{1, \cdots, n_{\inf}\}$ and write $E_{(1)} \geq E_{(2)} \geq \cdots \geq E_{(n_{\inf})}$ ranking $E_i, 1 \leq i \leq n_{\inf}$, in a decreasing order.

**Proposition 2.** If $E_{(n_{\inf})} > R_{\inf}E_{(1)^{1/2}} + (\chi^2_{\alpha}(1))/2$, then the Min-Unmask condition holds simultaneously for all the influential points $k \in S_{\inf}$.

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

To simplify the description, we introduce some notations. For any subset $U \subseteq \{1, \cdots, n\}$ with cardinality $n_U = |U|$ and any observation $Z_{k'}$ with $k' \in U$, we can draw at random with replacement subsets.
Proposition holds for group \( A_{r,U} \subseteq U \setminus \{ k' \} \), with the same cardinality \( n_{\text{sub},U} \), where \( n_{\text{sub},U} < n_U \). Similar to \( T_{\text{min},k} \), we define \( T_{\text{min}}(U, Z_k') = \min_{1 \leq r \leq m} n_{\text{sub},U}^r D_{r,U} k', U \), where \( D_{r,U} k' = p^{-1} || \hat{\rho}_{A_{r,U}} - \hat{\rho}_{A_{r,U} k'} ||^2 \). Denote by \( B_{r,U} \) the indices of non-influential observations in \( A_{r,U} \) and let \( O_{r,U} = A_{r,U} \setminus B_{r,U} \), \( 1 \leq r \leq m \). Let \( n_{\text{sub},U} \) be such that \( n_{\text{sub},U} = n_U k_{\text{sub},U} + 1 \). Then similar to \( F_{\text{min},k} \), we define \( F_{\text{min}}(U, Z_k') = \min_{1 \leq r \leq m} p^{-1} \sum_{t \in O_{r,U}} Y_t X_t^2 / (n_U k_{\text{sub},U}) \), which denotes the minimum of the joint effect of influential observations with indices in \( U \). And similar to \( F_{\text{max},k} \), one can define \( F_{\text{max}}(U, Z_k') \).

Obviously, when \( U = \{ 1, \cdots, n \} \), \( T_{\text{min}}(U, Z_k') \), \( F_{\text{min}}(U, Z_k') \) and \( F_{\text{max}}(U, Z_k') \) are exactly the same as \( T_{\text{min},k} \), \( F_{\text{min},k} \) and \( F_{\text{max},k} \), respectively.

Generally, suppose that \( E_{ij} \)'s can be separated into several groups in successive order, that is, \( G_j = \{ E_{(m_j-1)+1}, \cdots, E_{(m_j)} \} \), \( j = 1, \cdots, \tau \), such that \( 0 = m_0 < m_1 < \cdots < m_{\tau} = m_{\text{inf}} \). Denote \( I_j = \{(m_j-1) + 1, \cdots, (m_j) \} \), \( 1 \leq j \leq \tau \). Let \( M_0 = S_{\text{inf}} \), \( M_j = M_{j-1} \setminus I_j \) and \( U_j = M_{j-1} \cup S_{\text{inf}} \), \( 1 \leq j \leq \tau \). For simplicity, we assume that \( n_{\text{sub},U_j} \)'s are independent of \( j \), denoted still as \( n_{\text{sub},U} \), and that the sufficient condition in Proposition 2 holds for group \( G_j \), that is,

\[
E_{(m_j)}^1 > R_{\text{inf}} E_{(m_j-1)+1}^1 + (\chi^2_{1-\alpha}(1))^{1/2}, \quad 1 \leq j \leq \tau, \tag{3.2}
\]

which is referred to as \( \text{gMin-Unmask} \) Condition for simplicity. Then, similarly to the argument of Proposition 2 we see that \( \text{gMin-Unmask} \) Condition holds simultaneously for any \( Z_k, k \in I_j \) on the data set \( \{ Z_i, i \in U_j \} \), that is, \( E_{ij}^1 > F_{\text{max}}(U_j, Z_k)^1 + (\chi^2_{1-\alpha}(1))^{1/2} \). Consequently \( T_{\text{min}}(U_j, Z_k) \) with \( Z_k \in I_j \) will be large than \( \chi^2_{1-\alpha}(1) \) with high probability. If influential observations in \( I_1, \cdots, I_{\tau-1} \) are detected correctly and removed sequentially, the influential observations in group \( I_\tau \) can be detected successfully with high probability. We remark that the \( \text{gUnmask} \)-condition is much weaker than the condition in Proposition 2.

This motivates us to consider the following multi-round procedure. Define the set of influential observations identified in the \( j \)-th round as

\[
\hat{S}_{\text{min},j} = \{ k : P(\chi^2(1) > T_{\text{min}}(U_j, Z_k)) < q_k, Z_k \in U_j \},
\]

where \( q_k \) depends on the specific procedure used, similar to the discussion in Section 3.1, \( \hat{U}_j = \hat{U}_{j-1} \setminus \hat{S}_{\text{min},j-1} \) with \( \hat{U}_0 = \{ 1, \cdots, n \} \), and \( \hat{S}_{\text{inf},0} = \emptyset \). Finally, we can estimate \( S_{\text{inf}} \) by \( \hat{S}_{\tau'} = \cup_{j=1}^{\tau'} \hat{S}_{\text{min},j} \), where \( \tau' \) is such that \( \hat{S}_{\text{min},\tau'} = \emptyset \). Let \( F_{\text{PR}}(\hat{S}_{\tau'}) \) be the false positive rate associated with estimate \( \hat{S}_{\tau'} \).

**Proposition 3.** Suppose that \( (C1) \) holds and that FDR is controlled at level \( \alpha_0 \) in each round. Then \( E(F_{\text{PR}}(\hat{S}_{\tau'})) \leq \frac{\alpha_0}{1-\alpha_0} \).

Although the above iterative procedure can improve the performance of \( T_{\text{min},k} \) to overcome the masking effect, requiring only weaker \( \text{gMin-Unmask} \) Condition in (3.2), the computation of this procedure will be more costly if the number of rounds \( \tau' \) is large. On the other hand, the \( \text{gMin-Unmask} \) Condition will be easier to satisfy for larger \( \tau' \). Theoretically, \( \tau' \) can be as large as \( n_{\text{inf}} \), where \( \text{gMin-Unmask} \) Condition in (3.2) becomes \( F_{\text{min}} = \min_{1 \leq r \leq m} E_r > \chi^2_{1-\alpha}(1)/(1 - R_{\text{inf}})^2 \) by noting that \( E_{(m)} = E_{(m_{j-1}+1)} \), which is much weaker than the condition in Proposition 2. However, larger \( \tau' \) demands more intensive computing. If an early stopping strategy is adopted, it may still suffer from the masking effect.

As a quick summary, the test statistic \( T_{\text{max},k} \) is more efficient in dealing with the masking effect, because the strength of the influential observations required by \( T_{\text{max},k} \) in (ii) of Theorem 2 is much weaker than \( \text{gMin-Unmask} \) Condition (3.2) required by \( T_{\text{min},k} \), when \( m \) is large. Moreover, any procedure based on \( T_{\text{max},k} \) is computationally efficient, identifying the influential observations in just one round. However, \( T_{\text{max},k} \) may suffer from the swamping effect if the strong condition (i) of Theorem 2 is violated. On the other hand, the estimate \( \hat{S}_{\tau'} \) based on the statistic \( T_{\text{min},k} \) can maintain good FPR at the expense of more intensive computation. Taking advantages 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 propose the following algorithm to combine the strengths of the Max and Min statistics.

**Min-Max algorithm for estimating a clean set**
Initialization. Let $S_{\text{total}} = \{1, \cdots, n\}$ and fix $c = 1/2$. Repeat steps 1 and 2 until stop.

1. **Min-Step.** For the data indices in $S_{\text{total}}$, compute $\hat{M} = \{k : P(\chi^2(1) > T_{\min,k}) < \alpha_k, 1 \leq k \leq n\}$. 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 $S_{\text{total}} \rightarrow S_{\text{total}} \setminus \hat{M}$.

2. **Max-Step.** Estimate $\hat{S}_{\text{max}}$ as in Section 3.1 based on observations in $S_{\text{total}}$ and denote its complement $\hat{S}_{\text{max}}^c$ as an estimate of the clean set. If $|\hat{S}_{\text{max}}^c| \geq cn$, then stop; otherwise, go to Min-Step.

This algorithm identifies with high probability a clean dataset containing no influential points with cardinality at least $n/2$ by successively removing potential influential points. Here $\alpha_k$ 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 rational of this algorithm is, as argued, that the Max statistic $T_{\max,k}$ is aggressive in declaring influential while Min statistic $T_{\min,k}$ 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 $T_{\max,k}$ 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 large enough, we run the Min-Step again to remove further influential observations with strong strength. In our numerical study, we find that this algorithm is computationally very efficient, usually stops in 1 or 2 rounds.

With some abuse of notations, write $\hat{S}_c$ as the final clean set obtained by the Min-Max algorithm. Then its supplement, written as $\hat{S} = \{1, \cdots, n\} \setminus \hat{S}_c$, is an estimate of the set which contains all potential influential observations. However, $\hat{S}$ may still contain non-influential observations as the procedure for obtaining a clean set only aims to find a subset of the non-influential points. A further step to check whether any point in $\hat{S}$ is truly influential if necessary. This step, however, is easy since we have now a clean dataset. We now outline the exact procedure. For any $Z_i, i \in \hat{S}$, consider the data with indices in $\hat{S}_c$ and $\hat{S}_c^{(i)} = \hat{S}_c \cup \{i\}$, respectively. We then compute statistic $D_i$ as in Section 2 where $\hat{\rho}$ and $\hat{\rho}^{(i)}$ are computed on data set $\hat{S}_c$ and $\hat{S}_c^{(i)}$, respectively. Since $\hat{S}_c$ is a good estimate of the clean data containing no influential point, this leave-one-out approach will be effective for testing multiple null hypotheses in the form of $H_{0i}: Z_i$ is non-influential, $i \in \hat{S}$. If $\hat{S}_c$ is good, according to the results in HIM, $n^2 D_i$ will follow $\chi^2(1)$ distribution under $H_{0i}$ by Theorem 1 of [Zhao et al.] (2013), where $n_c = |\hat{S}_c| + 1$. The Benjamini-Hochberg procedure can then be applied to control FDR. Those whose corresponding hypotheses are rejected by the FDR procedure can be labeled as influential observations. The algorithm for detecting multiple influential observations, called Min-Max-Checking algorithm, is summarized as follows.

**Min-Max-Checking algorithm**

1. Estimate a clean subset $\hat{S}_c$ by the Min-Max algorithm;

2. Check for each $k \in \hat{S} = \{1, \cdots, n\} \setminus \hat{S}_c$ whether the $k$th observation is influential.

4 Simulation and Data Analysis

We evaluate the performance of MIP for detecting multiple influential points and compare it to HIM whenever possible. Throughout the simulation study, we set the sample size as $n = 100$ and the number of predictors as $p = 1000$. We generate $n$ observations from

$$ Y_i = X_i^\top \beta + \epsilon_i, \quad 1 \leq i \leq n, \quad (4.1) $$

where $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$, $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$. We then replace the first $n_{\text{inf}} = 10$ points in $(X_i, Y_i), i = 1, \cdots, n$ by $Z_{\text{inf}} = \{(X_{\text{inf}}^i, Y_{\text{inf}}^i), i = 1, \cdots, n_{\text{inf}}\}$ which are generated differently. The resulting dataset denoted as $Z_n$ thus may contain 10 influential points. For $\epsilon_i \sim N(0, 1)$ and $X_i \sim N(0, \Sigma)$ where $(\Sigma)_{ij} = 0.4^{|i-j|}$. The coefficient $\beta$ and how $Z_{\text{inf}}$ is generated are specified below.

We evaluate performance by assessing the success in identifying influential and non-influential points, the accuracy in estimating $\beta$ in Model (4.1), and the success in identifying the support of $\beta$. Let $S_{\text{inf}}$ be the index set of the influential points and $\hat{S}_{\text{inf}}$ as its estimate either by HIM or MIP. We first compute $TPR_{\text{inf}}$, the true positive rate for influential observation detection, and $FPR_{\text{inf}}$, the false positive rate for detection. That is, $TPR_{\text{inf}} = |\hat{S}_{\text{inf}} \cap S_{\text{inf}}|/n_{\text{inf}}$ and $FPR_{\text{inf}} = |\hat{S}_{\text{inf}} \cap S_{\text{inf}}|/(n - n_{\text{inf}})$. Denoting $FNR_{\text{inf}}$ as the false negative rate, we also compute the $F_1$-score defined as $F_1 = \frac{2TPR_{\text{inf}} \cdot FPR_{\text{inf}}}{TPR_{\text{inf}} + FPR_{\text{inf}} + FNR_{\text{inf}}}$. Obviously, the larger $F_1$, the better the corresponding method is.
Denote \( \hat{\beta} \) as an estimate of \( \beta \) which is based on the full data (FULL), or based on a reduced dataset after HIM is applied (HIM), or a reduced dataset after MIP is applied (MIP). In this paper, we estimate \( \beta \) via the Lasso. The accuracy of the estimation is evaluated by computing \( ERR = \| \hat{\beta} - \beta \| \) and we compare the accuracy of FULL, HIM and MIP.

Denote the support of \( \beta \) as \( \text{supp}(\beta) \) and its complement as \( \text{supp}(\beta)^c = \{1 \cdots p\} \setminus \text{supp}(\beta) \). We report the success in identifying the support of \( \beta \) by reporting

\[
\text{TPR}_{\text{vs}} = \frac{|\text{supp}(\beta) \cap \text{supp}(\hat{\beta})|}{|\text{supp}(\beta)|} \quad \text{and} \quad \text{FPR}_{\text{vs}} = \frac{|\text{supp}(\beta)^c \cap \text{supp}(\hat{\beta})|}{|\text{supp}(\beta)^c|}.
\]

In the following simulations, we set \( n_{\text{sub}} = n/2 + 1 \). That is, the random subsets \( A_r, r = 1, \cdots, m, \) all have cardinality \( n/2 \). We repeat each experiment 100 times and report the means of the quantities defined above. In implementing MIP, we set the number of random subsets as \( m = 100 \) for Example 2. For Example 1, we take \( m = 100, 200 \) or 300 to assess the effect of \( m \). In Table 2, because the \( \text{FPR}_{\text{inf}} \) of HIM can be large, we decided not to compute the coefficient estimates based on the reduced data to save space as long as \( \text{FPR}_{\text{inf}} > 0.7 \).

Finally, the FDR level is fixed at \( \alpha = 0.05 \).

### 4.1 Simulation setup

We simulate the data such that there exists a strong masking effect in Example 1 and a strong swapping effect in Example 2. Denote \( 0_s \) as a \( s \)-dimensional zero vector and \( 1_s \) as a \( s \)-dimensional vector of 1’s.

#### Example 1 (Strong masking effect).

We first generate \( n = 100 \) non-influential observations from \( [11] \) with \( \beta = (0.4, 0.5, 0.5, 0.6, 0.4, 0_{p-5})^\top \). Let \( i_0 = \arg \max_{1 \leq i \leq n} |Y_i| \). Then we replace the first \( n_{\text{inf}} = 10 \) non-influential observations by

\[
X_{ij}^\text{inf} = X_{nj} + I(j \in S_i) \cdot i/p, \quad Y_i^\text{inf} = Y_{i0} + \mu + \varepsilon_i^\text{inf} \cdot i/p, \quad 1 \leq j \leq p, 1 \leq i \leq n_{\text{inf}},
\]

where \( \{S_i\} \), with \( |S_i| = 10 \), are subsets of \( \{1, \cdots, 1000\} \) chosen independently with replacement, and \( \varepsilon_i^\text{inf} \sim N(0, 0.5) \). This example is designed such that the influential observations are clustered together and consequently many influential observations are masked by other influential ones. HIM based on leave-one-out will likely fail to identify many influential points. The simulation results are presented in Table 1 and plot (a) of Figure 2.

#### Example 2 (Strong swapping effect).

We set \( \beta = (0.2, 0.4, 0.5, 0.3, 0.2, 0_{p-5})^\top \) and generate influential observations according to the following scheme. Let \( w = (w_1, \cdots, w_{20})^\top \in \mathbb{R}^{20} \) with \( w_j = j \cdot 0.005\mu \). For \( i = 1, \cdots, n_{\text{inf}}, \) we let

\[
Y_i^\text{inf} = \text{sign}(\sigma_i) \cdot (\tilde{\beta}^\top X_{i}^\text{inf} + \varepsilon_i^\text{inf}), \quad X_{i}^\text{inf} \sim N(\nu_{\text{inf}}, I_p), \quad \nu_{\text{inf}} = (0_{900}, 0.5\mu 1_{100})^\top, \quad \tilde{\beta} = \beta + (0_{p-20}, w^\top)^\top,
\]

where \( \varepsilon_i^\text{inf} \sim N(0, 0.5) \) and \( \sigma_i \) is a binary variable with \( P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2 \) independent of \( (X_{i}^\text{inf}, \varepsilon_i^\text{inf}) \). For this example, when \( \mu \) is large, there exists a strong swapping effect. The simulation results are presented in Table 2 and plot (b) of Figure 2.

### 4.2 Summary of the simulation results

From Table 1 and Figure 2, we observe the following phenomena.

(1). The comparison between HIM and MIP when there exists a masking effect (Example 1) or a swapping effect (Example 2). From Table 1 and Figure 2, we see that HIM suffers from these effects seriously. For Example 1, we see that the \( \text{TPR}_{\text{inf}} \) of HIM is much smaller than that of MIP. Although its \( \text{TPR}_{\text{inf}} \) increases as \( \mu \) increases, the increment is slow and its \( \text{FPR}_{\text{inf}} \) increases at the same time. For Example 2, we see from Figure 2 that HIM works well when \( \mu \in [2, 4] \), but HIM suffers from the swapping effect when \( \mu \) is large, with its false positive rates much larger than 0.05.

On the other hand, MIP performs very well in Example 1 and 2. It is more powerful than HIM with larger \( \text{TPR}_{\text{inf}} \), while its \( \text{FPR}_{\text{inf}} \) is well controlled at the FDR level \( \alpha = 0.05 \). The price we pay is the computation cost, as \( m \) subsets are evaluated in MIP. Our simulation shows that the computing time of MIP increases
Table 1: Simulation results of Example 1 with different $\mu$.

| $\mu$ | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 6.5 | 7.0 |
|-------|-----|-----|-----|-----|-----|-----|-----|
| $TPR_{inf}$ | 0.780 | 0.820 | 0.940 | 0.960 | 1.000 | 1.000 | 1.000 |
| $FPR_{inf}$ | 0.003 | 0.005 | 0.004 | 0.003 | 0.003 | 0.002 | 0.002 |
| $F_1$ | 0.875 | 0.898 | 0.967 | 0.978 | 0.998 | 0.999 | 0.999 |
| MIP | $ERR$ | 0.570 | 0.568 | 0.553 | 0.518 | 0.525 | 0.502 | 0.507 |
| $m = 100$ | $TPR_{vs}$ | 0.944 | 0.944 | 0.944 | 0.964 | 0.936 | 0.972 | 0.960 |
| | $FPR_{vs}$ | 0.022 | 0.024 | 0.018 | 0.019 | 0.012 | 0.016 | 0.016 |
| $TPR_{inf}$ | 0.840 | 0.860 | 0.960 | 0.980 | 1.000 | 1.000 | 1.000 |
| $FPR_{inf}$ | 0.004 | 0.004 | 0.003 | 0.005 | 0.002 | 0.002 |
| $F_1$ | 0.911 | 0.923 | 0.978 | 0.988 | 0.997 | 0.999 | 0.999 |
| MIP | $ERR$ | 0.554 | 0.577 | 0.538 | 0.498 | 0.504 | 0.516 | 0.488 |
| $m = 200$ | $TPR_{vs}$ | 0.964 | 0.948 | 0.972 | 0.972 | 0.980 | 0.948 | 0.960 |
| | $FPR_{vs}$ | 0.021 | 0.020 | 0.015 | 0.012 | 0.019 | 0.015 | 0.012 |
| $TPR_{inf}$ | 0.860 | 0.920 | 0.960 | 0.980 | 1.000 | 1.000 | 1.000 |
| $FPR_{inf}$ | 0.003 | 0.004 | 0.003 | 0.002 | 0.003 | 0.006 |
| $F_1$ | 0.923 | 0.957 | 0.978 | 0.988 | 0.998 | 0.999 | 0.997 |
| MIP | $ERR$ | 0.587 | 0.529 | 0.529 | 0.540 | 0.523 | 0.479 | 0.488 |
| $m = 300$ | $TPR_{vs}$ | 0.956 | 0.976 | 0.964 | 0.956 | 0.956 | 0.972 | 0.976 |
| | $FPR_{vs}$ | 0.024 | 0.019 | 0.019 | 0.016 | 0.015 | 0.016 | 0.013 |
| $TPR_{inf}$ | 0.860 | 0.920 | 0.960 | 0.980 | 1.000 | 1.000 | 1.000 |
| $FPR_{inf}$ | 0.007 | 0.048 | 0.080 | 0.111 | 0.151 | 0.151 | 0.147 |
| $F_1$ | 0.972 | 0.421 | 0.388 | 0.331 | 0.571 | 0.535 | 0.607 |
| HIM | $ERR$ | 0.802 | 0.757 | 0.783 | 0.848 | 0.866 | 0.835 | 0.856 |
| $TPR_{vs}$ | 0.856 | 0.900 | 0.908 | 0.868 | 0.816 | 0.832 | 0.832 |
| | $FPR_{vs}$ | 0.040 | 0.040 | 0.043 | 0.044 | 0.033 | 0.038 | 0.036 |
| $ERR$ | 0.769 | 0.788 | 0.836 | 0.832 | 0.885 | 0.895 | 0.930 |
| FULL | $TPR_{vs}$ | 0.948 | 0.932 | 0.920 | 0.924 | 0.928 | 0.892 | 0.932 |
| | $FPR_{vs}$ | 0.047 | 0.051 | 0.055 | 0.052 | 0.055 | 0.056 | 0.061 |

Table 2: Simulation results of Example 2 with different $\mu$.

| $\mu$ | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|---|---|---|---|---|---|----|
| $TPR_{inf}$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $FPR_{inf}$ | 0.003 | 0.004 | 0.007 | 0.014 | 0.000 | 0.000 | 0.000 |
| $F_1$ | 0.998 | 0.998 | 0.996 | 0.993 | 0.999 | 1.000 | 0.999 |
| MIP | $ERR$ | 0.253 | 0.252 | 0.264 | 0.256 | 0.269 | 0.252 | 0.248 |
| $TPR_{vs}$ | 0.968 | 0.972 | 0.960 | 0.956 | 0.972 | 0.972 | 0.960 |
| | $FPR_{vs}$ | 0.014 | 0.015 | 0.016 | 0.012 | 0.020 | 0.016 | 0.014 |
| $TPR_{inf}$ | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| $FPR_{inf}$ | 0.018 | 0.093 | 0.357 | 0.502 | 0.859 | 0.989 | 1.000 |
| $F_1$ | 0.991 | 0.955 | 0.848 | 0.799 | 0.699 | 0.669 | 0.667 |
| HIM | $ERR$ | 0.263 | 0.305 | 0.442 | 0.490 | – | – | – |
| $TPR_{vs}$ | 0.968 | 0.924 | 0.696 | 0.684 | – | – | – |
| | $FPR_{vs}$ | 0.015 | 0.017 | 0.014 | 0.015 | – | – | – |
| $ERR$ | 0.738 | 0.884 | 0.914 | 1.011 | 1.072 | 1.162 | 1.383 |
| FULL | $TPR_{vs}$ | 0.188 | 0.072 | 0.032 | 0.016 | 0.000 | 0.004 | 0.004 |
| | $FPR_{vs}$ | 0.005 | 0.006 | 0.004 | 0.005 | 0.003 | 0.004 | 0.003 |
linearly with $m$. Therefore choosing a small or moderate $m$ can reduce the computing cost. Alternatively, by noting that subsets $A_1, \ldots, A_m$ are sampled independently, the computational time can be reduced if a parallel computing algorithm is used.

(2). From the comparison between the fit after MIP is used to remove influential points and the fit using the full data, it is clear that MIP is much better whenever there exist influential observations. In terms of variable selection, we see that the MIP based fits are slightly better than the HIM based fits and the FULL data based fits in Example 1. And in Example 2, the MIP based fits are much better. Now let us look at the effect of $m$. From Table 1 we see that MIP performs similarly for different values of $m$. Using $m = 300$ does not bring significant gain over $m = 100$. This shows that MIP may be insensitive to the choice of the number of the subsets.

(3). Finally, we compare MIP to the $\Theta$-IPOD method in She and Owen (2011). The simulation results using the latter for Example 1 and 2 are summarized in Table 3. Comparing Table 1 with Table 3 leads to the following conclusions. For Example 1, the true positive rates for identifying influential points are similar, but the false positive rates of $\Theta$-IPOD are much larger than those of MIP. For Example 2, the $TPR_{inf}$’s of the $\Theta$-IPOD method are much smaller than those of MIP for every setting, while its $FPR_{inf}$’s are much larger than MIP’s. We conclude that MIP is more effective than $\Theta$-IPOD. Part of the reason may be that the $\Theta$-IPOD method was developed based on a mean shift model, while our method does not assume the scheme of influentialness.

Table 3: Simulation results using the $\Theta$-IPOD method in [She and Owen (2011)]

| Example 1 | $\mu$ | $TPR_{inf}$ | $FPR_{inf}$ | $F_1$ |
|-----------|-------|-------------|-------------|------|
|           | 4.0   | 0.900       | 0.936       | 0.893|
|           | 4.5   | 0.980       | 0.980       | 0.922|
|           | 5.0   | 0.960       | 1.000       | 0.940|
|           | 5.5   | 1.000       | 1.000       | 0.947|
|           | 6.0   | 0.960       | 1.000       | 0.947|
|           | 6.5   | 1.000       | 1.000       | 0.928|
|           | 7.0   | 1.000       | 1.000       | 0.928|

| Example 2 | $\mu$ | $TPR_{inf}$ | $FPR_{inf}$ | $F_1$ |
|-----------|-------|-------------|-------------|------|
|           | 4.0   | 0.018       | 0.062       | 0.034|
|           | 5.0   | 0.092       | 0.116       | 0.113|
|           | 6.0   | 0.232       | 0.258       | 0.163|
|           | 6.5   | 0.258       | 0.382       | 0.202|
|           | 7.0   | 0.258       | 0.382       | 0.202|
|           | 8.0   | 0.232       | 0.382       | 0.202|
|           | 9.0   | 0.258       | 0.382       | 0.202|
|           | 10    | 0.258       | 0.382       | 0.202|

4.3 Real data analysis

As an illustration, we apply MIP to detect influential points in the microarray data from Chiang et al. (2006) which was previously analyzed by Zhao et al. (2013). For this dataset, we focus on 120 twelve-week-old male offspring that were selected for tissue harvesting from the eyes and for microarray analysis. The dataset contains over 31,042 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, 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 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 zero and standard deviation one (Huang et al., 2006). Applying Lasso to the full data using the default setting of glmnet function in R, we identify 15 significant variables and the $\ell_2$-norm of the estimated coefficient vector equals 0.097.

Applying HIM and MIP to this data with the FDR level at $\alpha = 0.05$, HIM finds 15 influential observations, while MIP obtains 7 influential observations. Interestingly, the set of influential points by MIP is a subset of that by HIM. In Figure 3, we plot the influential observations found by MIP in blue and the extra influential ones by HIM as red crosses, where the y-axis denotes the logarithm of the $p$-values obtained by using HIM as in plot (a) or using MIP as in plot (b). Note that, to make the plot more comparable, the checking step in the Min-Max-Checking algorithm is applied to all observations such that we can get a $p$-value for each observation. From this figure, we can see that the red crossed points identified by HIM as influential do not seem to have very small $p$-values.

To make further comparison, we use the ordinary least squares estimation on the important variables found via Lasso, after applying either HIM or MIP, to the non-influential point set identified by HIM. We compare their BIC score defined as $BIC = n \log(\text{RSS}/n) + k \log(n)$ where RSS is the residual sum of square, $n = 105$ is the same size after removing the 15 influential points identified by HIM, and $k$ is the number of variables used. Obviously, a model with a smaller BIC is preferred. Note $k = 9$ if HIM is used and $k = 6$ if MIP is.
applied. Because of the setup, this comparison favors HIM in some sense. It is found that BIC = −567.34 if HIM is applied for influential point detection and 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 further assess the performance of HIM and MIP, we artificially add influential points to the dataset and evaluate whether they can find these points afterwards. Specifically, we first remove the influential points 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

\[ X_{iS} = 1.1x_S + Z_S, \quad X_{iS'} = x_{S'}, \quad Y_i = 1.1y + \epsilon, \quad 1 \leq i \leq 10, \]

where \( Z \sim N(0, 0.01I_p) \), \( S \) is a random subset of \( \{1, \cdots, p\} \) consisting of 10 distinctive indices, \( Z_S \) is a subvector of \( Z \) with indices in \( S \), \( (x, y) \) is chosen randomly from 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 Benjamini-Hochberg procedure and repeat the process for 100 times. Then we compute the true positive rate (TPR) and false positive rate (FPR) of the two methods, respectively, for identifying these artificial influential points. It turns out that MIP gives a TPR of 1 and a FPR of 0.008, while HIM gives a TPR of 1 and a FPR as high as 0.585. Obviously, HIM suffers seriously from the swamping effect caused by the addition of new influential observations, while MIP does not seem to be affected by newly added observations.

5 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 proposed MIP framework can overcome the masking and swamping effects notoriously in influence diagnosis, and is able to identify multiple influential points with prespecified accuracy in terms of false discovery rate control.

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 using other quantities to define influence for example the generalized OLS estimator used for screening variables in Wang and Leng (2016).

Finally, we hope that this paper can bring 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 of the big data analytics, we believe that the issue of influence diagnosis will only become more relevant and hope that this paper can serve as a catalyst to stimulate more research in this area.

References

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

Atkinson, A. C. (1986). Masking Unmasked. *Biometrika, 73*, 533–541.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley, New York.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289–300.

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). *Proceedings of the National Academy of Sciences of the United States of America, 103*, 6287-6292.

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

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

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall, New York.

Draper, N. R. and Smith, H. (2014). *Applied regression analysis*. 3rd edition. John Wiley and Sons.

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *The Annals of Statistics, 42*, 324–351.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh-dimensional feature space. *Journal of the Royal Statistical Society, Series B, 70*, 849–911.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association, 88*, 1264–1272.

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

Huber, P. J. (2011). *Robust statistics*. Springer.

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics, 32*, 929–946.

Lawrence, A. J. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society, Series B, 57*, 181–189.

Nurunnabi, A. A. M. (2011). A diagnostic measure for influential observations in linear regression. *Communications in Statistics Theory and Methods, 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. *Journal of Applied Statistics, 41*, 1315–1331.

Pan, J., Fung, W., and Fang, K. (2000). Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference, 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.
Define $\hat{J}_r = N^{-2} \left\| \sum_{t \in B_r} \hat{Y}_t \hat{X}_t^\top \right\|^2$. Observe

$$J_{\max,n} \leq \max_{1 \leq r \leq m} | J_r - \hat{J}_r | + \max_{1 \leq r \leq m} \hat{J}_r.$$ 

The main idea of the proof is to show that the two terms on the righthand are small. For simplicity, we assume that each element of $(\hat{X}, Y)$ has population mean 0 and variance 1, that is, $\sigma_{xj} = \sigma_y = 1$ and $\mu_{xj} = \mu_y = 0, 1 \leq j \leq p$. Before the proof, we review some facts. For any $1 \leq t_1, t_2 \leq n$, define

$$K_{p,t_1,t_2} = p^{-1} \hat{X}_{t_1} \hat{X}_{t_2}, \quad K_{p,t,t} = p^{-1} \hat{X}_{t} \hat{X}_{t}^\top, \quad F_{t_1,t_2} = \hat{Y}_{t_1} \hat{Y}_{t_2}, \quad F_{t,t} = \hat{Y}_{t} \hat{Y}_{t}^\top.$$

Then by Lemma 1 of Zhao et al. (2013), we have $E(K_{p,t_1,t_2}) = 0$ if $t_1 \neq t_2$ and 1 if $t_1 = t_2$. Besides, $E(K_{p,t,t} - 1)^2 = O(p^{-2}l_p)$ and $E(K_{p,t_1,t_2})^2 = O(p^{-2}l_p)$, for any $t_1 \neq t_2$. In addition, $F_{tt} \sim \chi^2(1)$ due to $\hat{Y}_t \sim N(0, 1)$.

**Part I.** We show $\max_{1 \leq r \leq m} | J_r - \hat{J}_r | = O_p((\log(np)))(\log n)/((\log p)n^{-1/2})$.

**Step 1.** We first simplify the expression of $\hat{X}_t$ and $\hat{Y}_t$.

It is easy to see that for $1 \leq t \leq n$,

$$\hat{X}_{tj} = \hat{X}_{tj} \frac{\sigma_{xj}}{\sigma_{xj}} + \frac{\mu_{xj} - \hat{\mu}_{xj}}{\sigma_{xj}} \frac{\sigma_{xj}}{\sigma_{xj}} := \hat{X}_{tj}(1 + w_{x,nj}^{(1)}) + w_{x,nj}^{(2)}.$$
where \( w_{x,nj}^{(1)} = \frac{\sigma_{xj}}{\sigma_{xj} - 1} \) and \( w_{x,nj}^{(2)} = \frac{\mu_{xj} - \mu_{yj}}{\sigma_{xj} - \mu_{yj}} + \frac{\mu_{xj} - \mu_{yj}}{\sigma_{xj} - \mu_{yj}}(\sigma_{xj} - 1). \) Let \( w_{x,nj}^{(1)} = \max_{1 \leq j \leq p} |w_{x,nj}^{(1)}| \) and \( w_{x,nj}^{(2)} = \max_{1 \leq j \leq p} |w_{x,nj}^{(2)}|. \) By (C5) and simple calculation, we have, for some constant \( 0 < C < \infty, \)

\[
P(n^{1/2}w_{x,nj}^{(1)} > C \log p) \leq p^{-3}, \quad p(n^{1/2}w_{x,nj}^{(2)} > C \log p) \leq p^{-3}.
\]  

That is

\[
w_{x,nj}^{(1)} = O_p((\log p)n^{-1/2}), \quad w_{x,nj}^{(2)} = O_p((\log p)n^{-1/2}).
\]  

Similarly, let \( \hat{Y}_t = \sigma_y^{-1}(Y_t - \hat{\mu}_t), \) which follows standard normal \( N(0,1). \) Then

\[
\hat{Y}_t = \sigma_y^{-1}(Y_t - \hat{\mu}_y) = \hat{Y}_t + (\sigma_y^{-1}\sigma_y - 1)\hat{Y}_t + \hat{\sigma}^{-1}(\mu_y - \hat{\mu}_y)
\]  

\[
: = \hat{Y}_t + U_n\hat{Y}_t + \nu_n,
\]

where \( U_n \) and \( \nu_n \) are defined accordingly. Let \( w_y^{(1)} = U_n^2, w_y^{(2)} = \|\nu_n\|. \) Note that \( w_y^{(1)} = O_p(n^{-1}) \) according to the assumption on \( S_{R_n} \) in (C5). Similarly, we have \( w_y^{(2)} = O_p(n^{-1/2}) \) by (C5).

**Step 2.** Simplify the expression of \( \max_{1 \leq r \leq m} |J_r - J_r|. \)

Recall the definition of \( \hat{K}_{p,t_1,t_2}, K_{p,t_1,t_2}, \hat{F}_{p,t_1,t_2} \) and \( F_{p,t_1,t_2}. \) Define

\[
A_{t_1,t_2} = \hat{F}_{p,t_1,t_2}\hat{K}_{p,t_1,t_2} - F_{p,t_1,t_2}K_{p,t_1,t_2}, \quad 1 \leq t_1, t_2 \leq n.
\]

The we have

\[
|A_{t_1,t_2}| \leq |\hat{K}_{p,t_1,t_2}||\hat{F}_{p,t_1,t_2} - F_{p,t_1,t_2}| + |F_{p,t_1,t_2}||\hat{K}_{p,t_1,t_2} - K_{p,t_1,t_2}|.
\]

By Assumption (C1), we see that \( N_{B_r} > \delta n \) for all \( 1 \leq r \leq m, \) that is, \( N_{B_r} \) has the same order as \( n. \) By simple calculations, we have

\[
J_r - \hat{J}_r = N_{B_r}^{-2}\left\{ \sum_{t \in B_r} A_{tt} + \sum_{t_1 \neq t_2, t_1, t_2 \in B_r} A_{t_1,t_2} \right\}.
\]

Then it follows that

\[
\max_{1 \leq r \leq m} |J_r - \hat{J}_r| \leq \max_r N_{B_r}^{-1} \left( \sum_{t \in B_r} A_{tt} + \sum_{t_1 \neq t_2, t_1, t_2 \in B_r} A_{t_1,t_2} \right) \leq \max_r \left\{ \frac{N_{B_r}^{-1}}{1 \leq t \leq n} |A_{tt}| + \frac{N_{B_r}^{-1}}{t \neq t_2, 1 \leq t_1, t_2 \leq n} A_{t_1,t_2} \right\} \leq \max_r \left( \frac{N_{B_r}^{-1}}{1 \leq t \leq n} |A_{tt}| + \frac{1}{t \neq t_2, 1 \leq t_1, t_2 \leq n} A_{t_1,t_2} \right) \leq \max_r \left( \frac{N_{B_r}^{-1}}{1 \leq t \leq n} |A_{tt}| + \frac{1}{t \neq t_2, 1 \leq t_1, t_2 \leq n} A_{t_1,t_2} \right).
\]

**Step 3.** We study the terms in \( A_{t_1,t_2}. \)

**Step 3.1.** We show \( \max_{t_1,t_2} F_{t_1,t_2} = O_p(\log n) \) and

\[
\max_{t_1,t_2} |\hat{F}_{p,t_1,t_2} - F_{p,t_1,t_2}| = O_p((\log n)/\sqrt{n}).
\]

Because \( \hat{Y}_t \)’s are i.i.d. variables with distribution \( N(0,1), \|\hat{Y}_t\|^2 \sim \chi^2(1) \) and consequently, by the tail probability of \( \chi^2(1) \) distribution, we have \( \max_t |\hat{Y}_t|^2 = O_p(\log n). \) By Cauchy-Schwarz inequality, we see that \( \max_{t_1,t_2} F_{t_1,t_2} = O_p(\log n) \) holds. In addition, by the results in Step 1, applying Cauchy-Schwarz inequality and triangle inequality, we have

\[
\max_t |\hat{Y}_t - \hat{Y}_t|^2 \leq \max_t |U_n\hat{Y}_t + \nu_n|^2 \leq 2(\max_t |\hat{Y}_t|^2 w_y^{(1)} + (w_y^{(2)})^2) = O_p(\log n) + O_p(n^{-1}) = O_p((\log n)/n).
\]

(5.5)
Moreover, for any $1 \leq t_1, t_2 \leq n$, 
\[
\max_{t_1, t_2} |\tilde{K}_{p, t_1 t_2} - F_{p, t_1 t_2}| = \max_{t} |Y_{t_1}^T (\tilde{Y}_{t_2} - \hat{Y}_{t_2}) + (\tilde{Y}_{t_1} - \hat{Y}_{t_1})^T \tilde{Y}_{t_2} + (\tilde{Y}_{t_1} - \hat{Y}_{t_1})^T (\tilde{Y}_{t_2} - \hat{Y}_{t_2})| \\
\leq 2 \max_{t_1, t_2} |Y_{t_1}^T (\tilde{Y}_{t_2} - \hat{Y}_{t_2})| + \max_{t_1, t_2} |(\tilde{Y}_{t_1} - \hat{Y}_{t_1})^T (\tilde{Y}_{t_2} - \hat{Y}_{t_2})| \\
\leq 2 \max_{t_1, t_2} \|\tilde{Y}_{t_1}\| \max_{t_1, t_2} \|\tilde{Y}_{t_2} - \hat{Y}_{t_2}\| + \max_{t_1, t_2} \|\tilde{Y}_{t_1} - \hat{Y}_{t_1}\||^2 \\
= O_p((\log n)/\sqrt{n}).
\]  
(5.6)

**Step 3.2.** We show
\[
\max_{t_1, t_2} |\tilde{K}_{p, t_1 t_2} - K_{p, t_1 t_2}| = O_p((\log(p n))/\sqrt{n}).
\]

In fact, it is easy to see
\[
\max_{t_1, t_2} |\tilde{K}_{p, t_1 t_2} - K_{p, t_1 t_2}| = \max_{t_1, t_2} |p^{-1}[X_{t_1}^T (\tilde{X}_{t_2} - X_{t_2}) + (\tilde{X}_{t_1} - X_{t_1})^T X_{t_2} + (\tilde{X}_{t_1} - X_{t_1})^T (\tilde{X}_{t_2} - X_{t_2})]|.
\]  
(5.7)

For any $1 \leq t_1, t_2 \leq n$, we have
\[
p^{-1} \max_{1 \leq j \leq n} |X_{t_1}^T (\tilde{X}_{t_2} - X_{t_2})| \leq \max_{1 \leq j \leq n} |X_{t_1}^T X_{t_2}| w_{x,x}^{(1)} + \max_{1 \leq j \leq n} |X_{t_2}| w_{x,x}^{(2)}.
\]

Since $X_{t_2}$ are standard normal and $X_{t_1}$’s are independent with respect to $1 \leq t \leq n$, we have
\[
\max_{1 \leq j \leq n} |X_{t_1}^T X_{t_2}| = O_p((\log(p n))/\sqrt{n}),
\]
\[
\max_{1 \leq j \leq n} |X_{t_1}^T X_{t_2}| = O_p((\log(p n))/\sqrt{n}).
\]

Combining with (5.2) in Step 1, we have $p^{-1} \max_{t_1, t_2} |X_{t_1}^T (\tilde{X}_{t_2} - X_{t_2})| = O_p((\log(p n))/\sqrt{n})$. By similar arguments and noting $(\log p)/\sqrt{n} = o(1)$, we have
\[
p^{-1} |(\tilde{X}_{t_1} - X_{t_1})^T (\tilde{X}_{t_2} - X_{t_2})| \leq \max_{1 \leq j \leq p} |(\tilde{X}_{t_1,j} - X_{t_1,j})(\tilde{X}_{t_2,j} - X_{t_2,j})| \\
\leq \max_{1 \leq j \leq p} |\tilde{X}_{t_1,j}| w_{x,x}^{(1)} + w_{x,x}^{(2)} \\
max_{1 \leq j \leq p} |\tilde{X}_{t_2,j}| w_{x,x}^{(1)} + w_{x,x}^{(2)} \\
= O_p((\log(p n))/\sqrt{n}).
\]  
(5.8)

Therefore, we prove the conclusion on $\max_{t_1, t_2} |\tilde{K}_{p, t_1 t_2} - K_{p, t_1 t_2}|$.

**Step 3.3.** We show $\max_{t_1, t_2} |\tilde{K}_{p, t_1 t_2}| = O_p((\log(p n)))$.

Note
\[
\max_{t_1, t_2} |\tilde{K}_{p, t_1 t_2}| \leq \max_{t_1, t_2} |K_{p, t_1 t_2}| + \max_{t_1, t_2} |\tilde{K}_{p, t_1 t_2} - K_{p, t_1 t_2}|.
\]

The second term has been analyzed in Step 2. Consider the first term which satisfies
\[
\max_{t_1, t_2} |K_{p, t_1 t_2}| \leq \max_{t_1, t_2} |E(K_{p, t_1 t_2})| + \max_{t_1, t_2} |K_{p, t_1 t_2} - E(K_{p, t_1 t_2})|.
\]

Since $X_{t_1}$’s are standard normal, we have arguments similar to before that
\[
\max_{t_1, t_2} |K_{p, t_1 t_2} - E(K_{p, t_1 t_2})| \leq \max_{t_1, t_2} |X_{t_1}^T X_{t_2} - E(X_{t_1}^T X_{t_2})| = O_p((\log(p n))).
\]

For $E(K_{p, t_1 t_2})$, recall that $E(K_{p, t_1}) = 1$ and $E(K_{p, t_1 t_2}) = 0$ if $t_1 \neq t_2$. Thus, we have the conclusion of Step 3. Finally, combining all the results in Step 3, it follows that
\[
\max_{t_1, t_2} A_{t_1 t_2} = O_p((\log(p n)))(\log(n))(\log(p n))^{-1/2} = O_p((\log(p n))(\log(n))(\log(p)n)^{-1/2}).
\]

Combining with (5.4), we have the conclusion of Step 3 and it follows that
\[
\max_{1 \leq r \leq m} |J_r - \hat{J}_r| = O_p((\log(p n))(\log(n))(\log(p)n)^{-1/2}).
\]
This completes the proof of Part I.

**Part II.** We show the final conclusion by considering $\max_{r \leq m} J_r$. Note

$$J_r = N_{B_r}^{-2} \left[ \sum_{t_1 \in B_r} F_{t_1 t_1} K_{p, t_1 t_1} + \sum_{t_1, t_2 \in B_r, t_1 \neq t_2} F_{t_1 t_2} K_{p, t_1 t_2} \right].$$

Then

$$\max_r |J_r| = \max_r N_{B_r}^{-2} \left[ \sum_{t_1 \in B_r} |F_{t_1 t_1} K_{p, t_1 t_1}| + \sum_{t_1, t_2 \in B_r, t_1 \neq t_2} |F_{t_1 t_2} K_{p, t_1 t_2}| \right]$$

$$\leq \left[ \min_r N_{B_r} \right]^{-2} \left[ \sum_{1 \leq t_1 \leq n} |F_{t_1 t_1} K_{p, t_1 t_1}| + \sum_{1 \leq t_1, t_2 \leq n, t_1 \neq t_2} |F_{t_1 t_2} K_{p, t_1 t_2}| \right].$$

Note $N_{B_r} > \delta_1 n$. Then by Cauchy-Schwarz inequality, we have

$$T_1 := E\left[ \left( \min_r N_{B_r} \right)^{-2} \sum_{1 \leq t_1 \leq n} |F_{t_1 t_1} K_{p, t_1 t_1}| \right]$$

$$\leq (n \delta_1)^{-2} n E|F_{t_1 t_1} K_{p, t_1 t_1}|$$

$$\leq (n \delta_1)^{-2} n E(F_{t_1 t_1})^2 [\log(K_{p, t_1 t_1})]^{1/2}. \tag{5.9}$$

Noting that $F_{t_1} \sim \chi^2(1)$, we have that $E(F_{t_1 t_1})^2$ is bounded. Moreover, noting $E(K_{p, t_1}) = 1$ and $E(K_{p, t_1} - 1)^2 = O(p^{-2}l_p)$, we have

$$E(K_{p, t_1})^2 = E(1 + K_{p, t_1} - 1)^2 \leq 1 + E(K_{p, t_1} - 1)^2 = 1 + O_p(p^{-2}l_p).$$

Then $T_1 = O(n^{-1})$. Similarly, we have

$$T_2 := E\left\{ \left[ \min_r N_{B_r} \right]^{-2} \sum_{1 \leq t_1, t_2 \leq n, t_1 \neq t_2} |F_{t_1 t_2} K_{p, t_1 t_2}| \right\}$$

$$\leq (n \delta_1)^{-2} n(n-1) E|F_{t_1 t_2} K_{p, t_1 t_2}|$$

$$\leq \delta_1^2 [E(F_{t_1 t_2})^2 [\log(K_{p, t_1 t_2})]^{1/2}. \tag{5.10}$$

where $t_1 \neq t_2$ in the second inequality. By the Cauchy-Schwarz inequality, we have $E(F_{t_1 t_2})^2 \leq E(F_{t_1 t_1})^2 < \infty$. On the other hand, $E(K_{p, t_1 t_2})^2 = O(p^{-1}l_p/2)$. Therefore, $T_2 = O(p^{-1}l_p/2)$. Combining, we have $\max_{1 \leq r \leq m} |J_r| = O_p(p^{-1}l_p^{1/2})$. Finally combining the conclusions in Part I and Part II, we have

$$J_{\max, n} = \max_{1 \leq r \leq m} J_r = O_p \left( \log(n p) (\log(n) (\log p) n^{-1/2} + p^{-1} l_p^{1/2} \right).$$

**Proof of Theorem I**

Recall $n_{sub} = k_{sub} n$. Simple calculations shows that

$$D_{r, k} = p^{-1} \| \hat{\rho}_{A_{\cdot k}} - \hat{\rho}_A \|^2 = p^{-1} \left| \frac{1}{n_{sub}(n_{sub} - 1)} \sum_{t \neq k, t \in A_r} \hat{Y}_t \hat{X}_t^\top - \frac{1}{n_{sub}} \hat{Y}_k \hat{X}_k^\top \right|^2.$$

Consequently, it holds that

$$n_{sub}^2 D_{r, k} = p^{-1} \left| \frac{1}{n_{sub} - 1} \sum_{t \in B_r \setminus \{k\}} \hat{Y}_t \hat{X}_t^\top - \hat{Y}_k \hat{X}_k^\top \right|^2$$

$$:= p^{-1} \| W_{r, non} - \hat{Y}_k \hat{X}_k^\top \|^2.$$

By Lemma I we have $p^{-1} \max_{1 \leq r \leq m} \| W_{r, non} \|^2 = O_p(\zeta_{n, p} + p^{-1} l_p^{1/2})$. Therefore,

$$\max_{1 \leq r \leq m} n_{sub}^2 D_{r, k} = p^{-1} \| \hat{Y}_k \hat{X}_k^\top \|^2 (1 + O_p(\zeta_{n, p} + p^{-1} l_p^{1/2})).$$
On the other hand, based on assumption (C5) and the proof of Lemma 1 we have $p^{-1}||\hat{Y}_k\hat{X}_k^\top - \hat{Y}_k\hat{X}_k^\top||^2 \leq \max_{j,s}||\hat{Y}_{kj}\hat{X}_{kj} - \hat{Y}_{ks}\hat{X}_{kj}||^2 = o_p(1)$. That is, $p^{-1}||\hat{Y}_k\hat{X}_k^\top||^2 = p^{-1}||\hat{Y}_k\hat{X}_k^\top||^2(1 + o_p(1))$. Furthermore, note that $p^{-1}||\hat{X}_k||^2 = K_{p,t}\tau$ and that $E(K_{p,t\tau - 1})^2 = O(p^{-2}t^p)$. It follows that $p^{-1}||\hat{X}_k||^2 = O_p(1)$. Consequently, we have

$$p^{-1}||\hat{Y}_k\hat{X}_k^\top||^2 = ||\hat{Y}_k^2||^2 \left(p^{-1}||\hat{X}_k||^2 \right) = ||\hat{Y}_k||^2(1 + o_p(1)).$$

Note that $\hat{Y}_k$ follows $N(0,1)$. Therefore,

$$T_{\text{max},k} = \max_{1 \leq r \leq m} n_{\text{sub}}^2 D_{r,k} = ||\hat{Y}_k||^2(1 + o_p(1)) + o_p(1).$$

Consequently, $T_{\text{max},k} \rightarrow_d \chi^2(1)$. By nearly the same argument, it is easy to see that $T_{\text{min},k} \rightarrow_d \chi^2(1)$. This completes the proof. ■

**Proof of Theorem 2**

(1) We first prove the conclusion that $F_{\text{max},k} \leq R_{\text{inf},k}^2 d_{\text{inf}}$ mentioned just before Theorem 2. Note that $n_{\text{inf}} = n\delta_{\text{inf},n}$ and that $R_{\text{inf}} = \delta_{\text{inf},n}/k_{\text{sub}}$ where $k_{\text{sub}} > 0$ by assumption (C1). Denote $\hat{W}_{\text{inf},k,r} = n_{\text{inf}}^{-1} \sum_{t \in O_r} \hat{Y}_t\hat{X}_t^\top$.

Obviously we have $0 \leq |O_r| \leq n_{\text{inf}}$ due to the fact $O_r \subseteq S_{\text{inf}} \setminus \{k\}$. Recall the definition of $d_s$. Then

$$p^{-1} \max_{1 \leq r \leq m} ||\hat{W}_{\text{inf},k,r}||^2 \leq \max_{1 \leq r \leq m} \max_{t \in O_r} E_t \leq \max_{t \in S_{\text{inf}} \setminus \{k\}} E_t = d_{\text{sub}}(k).$$

Recall that $F_{\text{max},k} = p^{-1} \max_{1 \leq r \leq m} ||\hat{W}_{\text{inf},k,r}||^2$. Then, it holds that

$$F_{\text{max},k} \leq R_{\text{inf}}^2 \cdot \max_{1 \leq r \leq m} ||\hat{W}_{\text{inf},k,r}||^2 \leq R_{\text{inf}}^2 d_{\text{inf}}(k).$$

(2) We prove the conclusion of (i) and (ii). Recall that $n_{\text{sub}}^2 D_{r,k} = p^{-1}||W_{\text{non},k,r} + \hat{Y}_k\hat{X}_k^\top||^2$ by 3.1. By Lemma 1 it follows that $J_{\text{max},n} = p^{-1} \max_{1 \leq r \leq m} ||W_{\text{non},k,r}||^2 = O_p(\zeta_{n,p}\zeta_{n,p}^{-1/2}) = o_p(1)$. Consequently, by the Cauchy-Schwarz inequality, it holds that

$$T_{\text{min},k} = \min_{1 \leq r \leq m} n_{\text{sub}}^2 D_{r,k}$$

and

$$T_{\text{max},k} = \max_{1 \leq r \leq m} n_{\text{sub}}^2 D_{r,k}.$$

We prove the conclusion in (i). As $F_{\text{max},k} \rightarrow 0$, we have $T_{\text{min},k}$ and $T_{\text{min},k}$ converge in probability to $p^{-1}||\hat{Y}_k\hat{X}_k^\top||^2(1 + o_p(1))$. When $Z_k$ is non-influential, by the proof of Theorem 1 we have $p^{-1}||\hat{Y}_k\hat{X}_k^\top||^2 = E_k \rightarrow_d \chi^2(1)$.

We prove the conclusion in (ii). Due to the definition of $F_{\text{min},k}$, we can always find some $r_0 = r_0(m)$ such that $F_{\text{min},k} = p^{-1}||W_{\text{inf},k,r_0}||^2$. When $Z_k$ is influential, by 3.13 and the definition of $E_k$, it follows that

$$T_{\text{max},k}^{1/2} \geq \left[\max_{1 \leq r \leq m} n_{\text{sub}}^2 D_{r,k}\right]^{1/2} = p^{-1/2}||\hat{Y}_k\hat{X}_k^\top||^2(1 + o_p(1))^{1/2}$$

and

$$= (E_k^{1/2} - F_{\text{min},k}^{1/2})(1 + o_p(1))^{1/2}.$$

Since $E_k^{1/2} - F_{\text{min},k}^{1/2} > (\chi^2_{1-\alpha}(1))^{1/2}$, we have $P(T_{\text{max},k} > \chi^2_{1-\alpha}(1)) \rightarrow 1$. This completes the proof. ■
Proof of Proposition 1

Note that $J_r$ is defined for fixed $k$, that is, $J_r$ depends on $k$. Checking Step 2 of Part I and Part II in the proof of Lemma 1, we see that both $\max_{1 \leq r \leq m} |J_r - J_r|$ and $\max_{1 \leq r \leq m} |\tilde{J}_r|$ have upper bounds independent of $k$. Therefore, Lemma 1 actually holds uniformly over $k$, that is, $\max_k J_{max,n} = O_p(\xi_{n,p} + p^{-1/2}) = o_p(1)$.

By the proof of (ii) in the proof of Theorem 2, $T_{max,k}^{1/2} > (E_k^{1/2} - F_{min,k}^{1/2})(1 + o_p(1))^{1/2}$, where the term $o_p(1)$ depends on $\max_k J_{max,n}$ is independent of $k$. Therefore, $P(\cup_{k \in S_{inf}} \{T_{max,k}^{1/2} > E_k^{1/2} - F_{min,k}^{1/2}\}) \to 1$. Note that $\min_{k \in S_{inf}} T_{max,k}^{1/2} > \min_{k \in S_{inf}} E_k^{1/2} - \max_{k \in S_{inf}} F_{min,k}$. Since $\alpha_0$ is independent of $k$, we have $\max_{k \in S_{inf}} F_{min,k} < q_0^*$. Consequently, according to the assumption $E_k^{1/2} > (\chi^2_{-\alpha}(1))^{1/2} + a_0$, we have $P(\min_{k \in S_{inf}} T_{max,k}^{1/2} > (\chi^2_{-\alpha}(1))^{1/2}) \to 1$. Since $\chi^2(1)$ is the limit distribution under the null hypothesis of no influential observations, the $p$-values associated with observations of indices in set $S_{inf}$ are no more than $\alpha$ in probability. Therefore $\max_{k \in S_{inf}} p_{max,k} < \alpha$ with probability tending to 1.

Recall that $p_{max,i}$'s are the increasing order of $p$-value $p_{max,i}$'s. Let $k'$ be the largest $i$ such that $p_{max,i} \leq \alpha_0i/n$. The Benjamini-Hochberg procedure rejects hypothesis $H_{0i}$, where $1 \leq i \leq k'$. Denote by $[i]$ as the rank of $p_{max,i}$ in the series $p_{max,i}$'s. Let $\max_{i \in S_{inf}} [i]$ be the largest rank for $p_{max,i}, i \in S_{inf}$. If $\max_{i \in S_{inf}} [i]$ is less than $\alpha_0 [i]/n$ for $i \in S_{inf}$, then according to the rejection rule of the Benjamini-Hochberg procedure, all $H_{0i}$ with $i \in S_{inf}$ will be rejected. Noting that $\alpha = \alpha_0 \delta_{inf,n}$, we have in probability tending to one $\max_{i \in S_{inf}} [i] \leq \alpha_0 \delta_{inf,n} = \alpha_0 n_{inf}/n$.

On the other hand, it is easy to see that $\max_{i \in S_{inf}} [i] \geq n_{inf}$. Thus, it follows that $\max_{i \in S_{inf}} p_{max,i} \leq \max_{i \in S_{inf}} [i]/n$ in probability tending to 1. Therefore, all $H_{0i}$ with $i \in S_{inf}$ will be rejected by the Benjamini-Hochberg procedure.

Proof of Theorem 3 and Proposition 2

The proof of Theorem 3 is similar to that of Theorem 2. We first prove the conclusion in (i) of Theorem 3. By (5.12) and as $F_{min,k} \to 0$, we see that $T_{min,k} \to_k E_k$ and that $E_k \to_k \chi^2(1)$ for any $k \in S_{inf}$. Now we turn to conclusion (ii) of Theorem 3. Note that $\min_k T_{min,k}^{1/2} > (E_k^{1/2} - F_{max,k}^{1/2})(1 + O_p(\xi_{n,p} + p^{-1/2}))^{1/2} = (E_k^{1/2} - F_{max,k}^{1/2})(1 + o_p(1))^{1/2}$. According to the argument in the proof of Proposition 1, the term $O_p(\xi_{n,p} + p^{-1/2})$ is independent of $k \in S_{inf}$. Therefore, $P(T_{min,k}^{1/2} > E_k^{1/2} - F_{max,k}^{1/2}) \to 1$. Combining with the assumption $E_k^{1/2} > F_{max,k}^{1/2} + (\chi^2_{-\alpha}(1))^{1/2}$, we have the conclusion as desired.

Finally, we prove Proposition 2. Recall that $F_{max,k} \leq B_{inf}^2 d_{inf} \{k\}$ in (5.11). The sufficient condition in Proposition 2 is derived from the fact that $d_{inf} \{k\} \leq E(1)$ and the Min-Unmask condition of Theorem 3.

Proof of Proposition 3

We consider only the case when $K = 2$. The proof of the general case is similar. Denote by $n_1$ and $n_2$ as the expected number of hypothesis rejected in round 1 and 2, respectively. Since the FDR level is controlled at $\alpha_0$ in each round, then for estimate $S_{min}^1 \cup S_{min}^2$, the expected number of falsely rejected hypotheses is less than $\alpha_0(n_1 + n_2)$ where $n_1 + n_2$ is the expectation of the total number of rejected ones. Therefore FDR is still controlled at level $\alpha_0$, that is, $\tilde{R}_{non}/(\tilde{R}_{non} + \tilde{R}_{inf}) \leq \alpha_0$, where $\tilde{R}_{non}$ is the expected number of non-influential observations that are falsely labeled as influential ones, and $\tilde{R}_{inf}$ is the expected number of influential observations that are correctly identified. Due to the fact $\tilde{R}_{inf} \leq n \delta_{inf,n}$, we have $\tilde{R}_{non} \leq \alpha_0(1 - \alpha_0)^{-1} n \delta_{inf,n}$. Then

$$E(\text{FPR}(\tilde{S})) = \frac{\tilde{R}_{non}}{n(1 - \delta_{inf,n})} \leq \frac{\alpha_0 \delta_{inf,n}}{(1 - \alpha_0)(1 - \delta_{inf,n})} \leq \frac{\alpha_0}{1 - \alpha_0},$$

where we use in the last equality the assumption that $\delta_{inf,n} < 1/2$ in (C1).